2013

Improving Qualitative Assessment In Higher Education

Chad Woodson Russell
University of Mississippi

IMPROVING QUALITATIVE ASSESSMENT IN HIGHER EDUCATION

A Dissertation
presented in partial fulfillment of requirements
for the degree of Doctor of Philosophy
in the Department of Leadership and Counselor Education
The University of Mississippi

by

CHAD W. RUSSELL

May 2013

ABSTRACT

This dissertation considers in turn the role education plays in civil, democratic society; the role assessment plays in education; and the role theoretical constructs and cultural contexts play in assessment. Then, through literature review, document analysis, and interviews, the analysis investigates, identifies, and recommends grounded-theory-derived practices for improving qualitative assessment in higher education settings. The process of qualitative assessment is understood as being heuristic and continual, requiring re-examination and revision to maintain both its validity and reliability. To this end, rubrics are essential to efficiently and reliably assessing everything qualitative, whereas the realities of institutional culture and politics require adroit leadership from educators and administrators, drawing from manifest praxes in organizational theory, management theory, and political theory, to affect progressive change.

DEDICATION

To my parents, Bill and Jeannice, for doing all the things that parents need to do.

ACKNOWLEDGEMENTS

I wish to extend my thanks to my various interviewees and correspondents for their particular insights: Shanna Flaschka, Dr. Paul Hill, Dr. Alice Myatt, Kathleen Schmidt, Laura Schrock, and Dr. Sue Smith. I also wish to extend my thanks to the members of my committee for their helpful inputs and guidance: Dr. RoSusan D. Bartee, Dr. Joe M. Blackbourn, Dr. Douglas R. Davis, and especially Dr. Lori A. Wolff, committee chair and my academic advisor.

TABLE OF CONTENTS

Abstract ii

Dedication iii

Acknowledgements iv

Chapter 1: Introduction 1

Chapter 2: Review of the Literature 9

Chapter 3: Methodology 23

Chapter 4: Studies in Assessment Praxis and Essential Qualifying Technology 39

Chapter 5: Additional Considerations, General Conclusions, and Specific Recommendations 101

List of References 113

List of Appendices 127

Appendix A: Information Sheet and Consent Form 128

Appendix B: Protocol for Interview Questions 131

Appendix C: Perelman Essay 135

Vita 139

CHAPTER 1: INTRODUCTION

Statement of the Problem

Assessing learning outcomes is an essential part of establishing both the accountability for, and the credibility of, the credentialing function of higher education, and as such, evaluations of instructional effectiveness play a central role in accreditation. From modern *critical theory*, higher education in turn may be seen to play a central role in establishing and maintaining democracy and social justice within civilization (Giroux, 1997). The methodology of learning assessment shapes the curriculum and pedagogy of education, and thereby shapes — indirectly — the knowledge base, worldview, and values of society at large (Eisenhower, 1961; Fulbright, 1970; Giroux, 2005). Therefore, the form and function of learning assessment has significant consequences for the human condition: assessment shapes curriculum, curriculum shapes education in turn, and finally education shapes society. Thus, what is emphasized in learning assessment — both in terms of process and outcomes — becomes, for better or worse, what is emphasized in social structures. These social structures then form a *feedback loop* reinforcing the trends in higher education that give rise to them, and, in the present cultural climate, this cycle may be *considered harmful* inasmuch as it leads to *regressive* tendencies in both education and society (Giroux, 2007).

Since the advent of modern statistical techniques and electronic computers to efficiently utilize those techniques to analyze data and subsequently mine the results for further analysis, *standardized*, *normalized*, and purely-quantitative methodologies of learning assessment have

become predominant in educational practice, with non-trivial social consequences (Giroux, 2007). Qualitative assessment methodologies have not enjoyed similar advances in computational efficiency, and in comparison to quantitative methodologies, remain both more *subjective* and more labor-intensive for all participants. This study is intended to identify qualitative assessment strategies and techniques that have either shown promise toward or have been demonstrably successful in rectifying this quantitative/qualitative discrepancy of *efficiency*, *validity*, and *reliability*, and thereby provide recommendations to offset the imbalance between emphasis on qualitative versus quantitative assessments that is found in current practice in higher education. By returning meaning and context to a more-central role in learning assessment, a renewed emphasis upon qualitative perspectives may address many of the perceived shortcomings that higher education has drawn frequent criticism over in recent years (Hersh & Merrow, 2005).

Purpose Statement

The purpose of this grounded-theory, multiple-case study is to identify and examine best practices in qualitative learning assessment in higher education so that generalizable methodological constructs comparable to validity and reliability in quantitative assessments may be identified, fostering the transferability of virtuous qualitative assessment techniques to other postsecondary institutions and their accreditation processes.

This investigation, being grounded in particular contemporary social and educational theories that drive the need for this research and shape its initial expectations, takes the form of selected in-depth case studies of programs and institutions recognized as being successful at qualitative assessment and residing within those theoretical contexts. This study does not attempt to investigate the theoretically-assumed broader social consequences of successful programs in

qualitative learning assessment, but does make recommendations for how similar program effectiveness might be realized at other institutions using either existing programs or evolutions of them and relying upon the theoretical constructs that subsequently emerge from the research.

Theoretical Contexts and Definitions

The grounding of this study is based upon the synergistic interaction of two theoretical perspectives, *critical theory* and *constructivist learning*, as they may guide qualitative assessment and as education functions within the present-day socio-political and cultural climate of the United States.

Critical theory, in its sociological application, is concerned with the dynamics of social power and the belief systems that enable and result in that power (Giroux, 1990). In particular, postmodern critical theory examines the relationships between authority and injustice that arise from capitalism as an economic system, and by extension, a political system. One of the central ideas of recent critical theory is *commodification* — the reduction of all values to market-based ones, and how this reductionism is anti-humanistic. Viewing all human activity through the lens of the marketplace has dehumanizing consequences — as it considers only extrinsic value at the expense of intrinsic value — and historically, moral and ethical systems have arisen (or evolved) to counter — or at least limit — such *objectifying* perspectives of judgment (Giroux).

Unfortunately for the public welfare, the social and political climate in recent generations has been subject to a feedback loop of materialism wherein social and cultural capital has become subjugated to purely economic capital, and the education system has been both a victim of, and complicit in, this process (Giroux, 1983). Although the causes of this materialist, anti-humanist shift may be complex and varied, the necessary participation in and contributions to it by traditionally-autonomous institutions of higher education is by no means necessary nor

irreversible. Indeed, it has been noted (Fulbright, 1970) that education is a public good which should not cede its autonomy to any government; this is the reason, for example, why the U. S. Constitution (at least in its original adopted and uninterpreted form) does not grant the federal government any powers over education: public education is intended to serve as a check against political abuses of power, governmental or otherwise. It is for this reason, again for example, that American colleges and universities are accredited not by the federal government, but by relatively-autonomous agencies they themselves have established for the purpose.

To this end, recent initiatives and directives from the federal government exclusively emphasizing quantitative assessments of learning have been viewed as harmful by many educators not only because such analyses drive normalization practices that are not respective of the dignity and worth of individual students *qua* individuals (Hopmann, 2008), but because standardized testing paradigms have been viewed as corrupting instruction away from progressive student-centered models of educational assessment and regressively back toward discipline-centered ones (Broadhead, 2002). Furthermore, there are systematic issues raised from the reliance upon quantitative measures to inform policy, inasmuch as anything in the social sciences that is measured for purposes of control becomes, in response to the exercise of control through it, itself malleable and less-reliable over time (Goodhart, 1975).

The concern among progressive educators and social critics is that students are becoming merely well-trained workers *in lieu* of becoming critical-thinking citizens (who are also capable of doing work) (Asher, 2009). Industry often decries the lack of critical thinking skills in the modern workforce, yet it is industry itself that may be seen to drive much of the current trend away from critical thinking in the pedagogy in favor of standardized — and thus, typically quantitative — performance testing in learning. Increased calls from outside of academia for

accountability in higher education must be met by higher education itself, lest other agencies take it upon themselves to compel accountability in ways of their own choosing (Griffiths, Vidovich, & Chapman, 2008). Thus, studies such as this one may provide useful insight into problems many inside higher education might not yet realize we will soon face.

As to education theory specifically, constructivist models of learning emphasize the emergence of knowledge *within* the learner, rather than its mere transmission *to* the learner. Among contemporary educational theorists, such models of education — based upon imparting the learner-centric *skill* of seeking and acquiring knowledge rather than a discipline-centric *database* of mere facts — are seen as being more relevant to the increasing complexities of the modern world (Nussbaum, 2005). These newer, more modern models, however, are more difficult to assess, as they tend to rely upon subtle processes of information creation and synthesis within the learner rather than straightforward accumulation and accommodation of ostensible, sanctioned facts. Creativity, as a particular measure of learning, is not well-suited to quantitative analysis, as both implicitly and by practical definition *originality* may not be standardized (Baker, 2004).

Thus there are both individualized reasons, such as personal excellence and development, and contextualized reasons, such as the well-being of society at large, to pursue qualitative forms of learning assessment, but practical opposition from either pragmatic or political forces has been problematic to that end (Carless, 2005; Nkosana, 2008). The likely-unsolvable conundrum confronts us: how does one, as an educator, facilitate and then meaningfully assess a student's critical thinking skill using a standardized multiple-choice test? In philosophical practice, the resistance of a question to being answered may often be taken as an indicator that the question is not well-formed. This study attempts to recapitulate the issue of assessment in such a way that

educators and policymakers — as well as other stakeholders — may, from the examples and analyses of emergent theoretical constructs, find helpful methods of qualitatively assessing learning in reliable and repeatable ways that avoid the negative consequences of purely-quantitative approaches without introducing negative consequences of their own.

Delimitations

As an example of how subjective judgments may be standardized in qualitative assessments, *rubrics* linked to *learning outcomes* are already a familiar tool for modern educators. The existence of mechanisms such as rubrics suggests that this present research inquiry has promising precedent: *criteria*, *levels* of achievement, and assorted *descriptors* within those levels all stand as methodological paradigms, upon which those rubrics are built, limiting the subjectivity of the assessor. Typically, this subjectivity is further limited by the assessor making the rubric available not only to the students whose work will be assessed under it, but also to the overseeing institutional accreditation agency. Accreditors may subsequently compare and contrast rubrics within and across departments, programs, and institutions pursuant to institutional assessment processes. Some instructors, under the aegis of academic freedom, may demonstrate some resistance to committing their judgment processes to paper in the manner of a rubric, but such habits may be merely reflective of the perpetual tension between academic freedom and public accountability, and that issue is not germane to this study. What is relevant to this study is that there is already an explicit acknowledgement among educators that consistent standards of qualitative assessment are indeed possible, even if the process is still in its early developmental stages in academe (Carless, 2005; Hopmann, 2008).

In addition, there is a recognized tradition in higher education of *writing across the disciplines* as having value in developing critical thinking skills — the rationale being that

showing the ability to express an idea in words demonstrates not only mastery of the intellectual content of the idea itself, but also the successful cognitive integration of that idea with other concepts in the learner's worldview (Herrington & Moran, 1992). The problem with assessing writing, however, is the significant time and labor required — first in the generation of the text by the student, and then in the evaluation of the text by the instructor. More complex instruments such as portfolios or live task performances may be even more labor-intensive and time-consuming.

Preliminary Research Questions

Two avenues of inquiry suggested themselves as starting points for the investigation: 1) how can the place and function of rubrics in curriculum and pedagogy be examined to yield practical utilities and theoretical constructs that transcend discipline silos, unique institutional cultures, and the individual assessors using them, and 2) what practices — technological or ideological — exist or may be developed to efficiently manage the workload of both students and instructors engaged in qualitative assessments? Further investigation naturally gives rise to questions and insights beyond these, as detailed in the methodology and particular case studies below.

Structure of the Study

This qualitative study is organized conventionally. After this introductory chapter, a brief literature review chapter is presented. The literature review is not intended to be completely exhaustive, but provides a broad and diverse survey examining in more detail the various factors providing the background and context behind the inquiry. Following the literature review, Chapter 3 explains the study's methodology in all its assorted aspects, addressing and discussing the many considerations relevant to a qualitative investigation and the case study format, and

how this study operates within them from the researcher's perspective. Subsequent chapters provide the individual, detailed case studies of programs and institutions purposefully selected for examination. The final, summary chapter discusses the findings and recommendations of the study within the context of the issues raised in both this introductory chapter and the individual case study chapters, again from the researcher's perspective.

CHAPTER 2: REVIEW OF THE LITERATURE

Introduction

This chapter presents a summary of the literature examining the consequences of, methods for, and purposes behind recent changes in the policy and theory of liberal, public education as the foundation of a democratic society. In this context, constructivist educational theory is a long-standing area of research interest in the philosophy of education dating at least as far back as the early 20th Century and the work of John Dewey, but its roots may be seen in the writings of 19th Century thinkers such as Friedrich Nietzsche and even 18th Century thinkers such as Thomas Jefferson. This review summarizes both the theoretical dynamic within which education may be viewed as an ontological-sociological technology as well as the unavoidable political elements which arise from the structure of that dynamic and to which education is thus subjected, for better or worse, in the contemporary cultural environment. This historical and philosophical framework may then form the basis of the case studies and "best practices" analyses and recommendations that follow.

The review of literature is presented as follows. First, an historical overview of the controversy surrounding the publication of *The Bell Curve* is summarized, to provide a background for the resulting problem of dehumanizing standardization. Second, the technology of education from the theories of John Dewey is examined in its historical context to provide the humanistic foundation for modern educational theory and educational reform. Third, from the more-recent writings of Michel Foucault, the concept of the learning apparatus is analyzed and

developed for application to the problems modern, progressive educational theory faces. Fourth, using recent and less-recent research, the architecture of control is discussed for both an explanation of its evolution and function, and to show how the goals of education were shifted in the middle of the last century away from their traditionally humanistic purpose. Fifth, the dehumanizing consequences of overly-quantified assessment practices and perspectives are examined and considered harmful, as foreseen by some researchers.

The Origin and Context of the Problem: Social Darwinism and Sorting

The publication, and widespread exposure, of *The Bell Curve* text (Herrnstein & Murray, 1994) has established itself as a watershed event in the history of educational theory, for reasons its authors may well have intended (Fendler & Muzaffar, 2008). The Gaussian distribution — the shape of which has often been interpreted as bell-like — is a mathematical model of probability with its own well-established history as a data-derived phenomenon in both the physical and behavioral sciences (Stigler, 1986). However, this model has been extended, largely due to the influence of *The Bell Curve* on public policy decision-making, from a descriptive model resultant from objective analysis into a subjective, normative model used — or arguably misused — for prediction and the accumulation and exercise of social power and control (Fendler & Muzaffar; Steele & Aronson, 1995).

The "bell curve" is a form of quantitative assessment; in the original presentation, Herrnstein and Murray (1994) specifically applied the model to intelligence quotient (IQ) testing, and found significant correlations between IQ and race and socio-economic status. The authors then used these correlations to — inappropriately without controlled experimentation — not merely hypothesize, but genuinely posit, the existence of a causal agency, which they identified as being genetic, and therefore racial, in origin (Fendler & Muzaffar, 2008). Although the

explicit racism of this methodologically-questionable finding drew much subsequent criticism and defense of its particular "biological determinism" conclusions (Gould, 1996; Murray, 1995), the general appropriateness of normative, quantitative assessment practices in the behavioral sciences became essentially unassailable as a result of the debate. This outcome was because although the particular application of statistical modeling to IQ distribution in the general population was subsequently examined for flaws in the research design, the validity of the mathematical technique was generally accepted by both sides of the debate (Gardner, 1995). Such debate, which continues through the present day, centers on the reflexive and subjective nature of the reified theoretical constructs of "race" (Smedley, 1998) and "intelligence" (Neisser & Boodoo, 1996), and thus their ultimate meaninglessness and uselessness as objective-analytical tools. The validity of such analysis overall when applied to the inherently at-least-partially-subjective-and-therefore-arguably-non-scientific phenomenon of human behavior, however, has not been examined with as much rigor or vigor (Plucker, 2003). The scientific method has certainly established its utility in advancing the understanding of human behavior, but given the aforementioned subjectivity, it is naïve — even willfully so — to presuppose quantitative methods alone will be sufficient to the task. Scientific methods are only as good as the epistemologies and ontologies upon which they are erected, and it is necessary to call upon the qualitative from time to time to give meaning to measurement.

For the present consideration, the point is that the application of normal probability distribution (NPD) assumptions to quantitative assessments at all levels of education is ultimately a regressive, rather than a progressive, process (Fendler & Muzaffar, 2008), and this process is often deliberately intended to advance a particular political agenda regarding society (Murray, 2007). Furthermore, there is a methodological and logical fallacy involved in the

circular reasoning behind the numerical definition of a normalized standard to fit to observed data and then the assertion that the measurement of that standard represents some instance of an actual phenomenon naturally occurring in the population; there is no such thing in the field as an "average" person, since "average" is a mathematical construct, not a natural entity. Therefore, applying the bell curve model to outcome assessment — learning assessment in particular — generates a phenomenological ontology regarding an emergent entity who is conceptualized as the Average Student (Fendler & Muzaffar). This categorization, and the sorting schema it subsequently inspires, does not necessarily represent an actual, natural state of affairs in the population, but is merely a self-serving artifact of the way the analysis is structured. Although the resulting tendency toward sorting within the educational system has often been considered unwelcome, arguments against sorting have remained historically ineffective, primarily due to the influence of politics on the policy process (Hacking, 1990, 2002).

The Technology of Education

The notion of progressive education has its origins in the early 20th Century writings of pragmatist philosopher and philosopher of education John Dewey. The theory emerges as an extension of the traditional notion of a liberal education (as promoted by Thomas Jefferson, for example) into a constructivist model wherein both the individual and society develop in a mutually-reinforcing teleological process of increasing complexity and excellence, and hence, "progress" (Dewey, 1900, 1916). In such a scenario, the interdependent thriving of both the society and the individual operates in a scientific and moral way — a "technology" — to insure the well-being of both; this synergistic coupling represents the ultimate expression of democracy, as idealized by the ancient Greek philosopher Aristotle.

Central to this growth process is the emphasis that should be placed upon a student's problem-solving skills; in addition to the basic social skills of learning to live and work cooperatively with others, schools should place the emphasis on developing the student's judgment rather than mere knowledge, as this is today, perhaps even more so than a century ago, the key skill individuals need to flourish in an increasingly-complex society. It is Dewey's position that assessment should be based on an individualized metric (judgment) rather than an objective, normalized one (knowledge), and little well-supported rebuttal of this point — much less refutation — has been forthcoming in the decades since it was first made.

The Apparatus of Learning

Learning itself constitutes a form of governance. Historically, modern educational theory examines the two environmental contexts within which the learner learns (Lewis, 2007). After Dewey, the process of learning can be viewed as occurring in a state of tension between the developing autonomy and self-realization of the student, and the authority of family and the state that impose formal education upon the pupil. This dichotomy recapitulates, on the personal scale, the larger power dynamic of (all) society wherein a dialectic between individual autonomy and public authority is constantly playing out. The writings of Michel Foucault (1972, 1994) explore and develop this general sociological process in much more detail; the key conclusion is that standardization — whether it is in curriculum, or pedagogy, or assessment — represents an attempt to assert sovereign control through the mechanism of instruction. At times, this process may even be characterized as a form of indoctrination, especially in circumstances where a particular educational system is designed and implemented to strengthen authority at the expense of autonomy (Mason, 2008; Piro, 2008).

In contrast, modern educational principles such as “life-long learning” and “critical thinking” are considered valuable tools — and in combination, an apparatus — for increasing the power and autonomy of the individual as a member of (an implicitly democratic) society (Lewis, 2007; Simons & Masschelein, 2008). This individual empowerment may in turn form the basis of a general social empowerment — after Foucault — which may then shift authority away from oligarchic models of governance and onto the populace directly through the accumulation of socio-economic capital. Most saliently, this capital is generated by the learning process itself, rather than any particular learning outcome, and is therefore more qualitative than quantitative in nature. Naturally, authoritarian governments seek to inhibit or prevent the accumulation of much of this capital by the public in order to preserve the concentration of power in the structures of governmental authority and thereby perpetuate and increase that authority.

As suggested above, the best and most-confirmed path to the manifestation of an Aristotelian/Jeffersonian utopian democracy is found in the technology of education set forth by John Dewey (Margonis, 2009). The inculcation of autonomous values, rather than the indoctrination of authoritarian ones, is the defining, crucial “technology” involved in progressive education, and therefore the growth and development, of the individual learner. This process then fosters the principle of self-governance under which both the individual student and the “multitude” of a public comprised of such individuals operate. Ultimately, education, and the form its practical components and practices take, may be conceived of as the defining and controlling factor in what form of governance the society operationalizes overall. In as many words, education is governance — of both the individual and the society (Piro, 2008). Control of education ultimately yields control of government, and, more crucially, *vice-versa*.

The Architecture of Control

The natural mechanisms of governance drive governing systems toward authoritarianism (Helfenbein & Shudak, 2009), and contemporary developments in information/instructional technology and *realpolitik* are driving new challenges to progressive education as authoritarian control becomes easier to exercise and more generally-accepted politically (Masschelein, 2004). History demonstrates that the consolidation of power drives the further consolidation of power (Nietzsche, 1887); this is the reason that the dispersal and sharing of power as widely as possible — through some democratic system of government — is considered the best defense against tyranny. Although fully-democratic societies have their own problems of "groupthink" and its like, they are still considered preferable to the regressive dehumanization that inevitably results from tyranny placing the preservation of state power and control ahead of individual flourishing (Piro, 2008).

Conversely, for government — even and especially a progressive, democratic one — the main challenge in the exercise of its authority is striking a balance between individual and general welfare (Simons & Masschelein, 2006, 2008). This is further complicated by uncertainties among theorists within the education profession as to how best to address these issues (Masschelein & Quaghebeur, 2005), or even if they should be addressed in the current political climate (Depaepe, 2007).

Another obstacle to progressive educational reform in service to the advancement of democratic governance lies in the fact that systematic, methodical attempts to promote democracy and democratic principles in curriculum and pedagogy are largely ineffective in the face of economic pressures and institutional (government) power as individuals — particularly as

capitalism has spread around the world — become preoccupied with fundamental, personal economic needs in place of broader cultural ones (Helfenbein & Shudak, 2009).

Together, all these factors combine to create an already-significant advantage for institutional power to maintain its authority and control, but a new mechanism has been deployed to further enhance that advantage: ideology shifts through language shifts (Helfenbein & Shudak, 2009). Contemporary developments in mass-media technology have allowed public discourse to be controlled and shaped in historically-unprecedented ways, and as a consequence, the ontologies that individuals devise to understand the modern world are heavily-influenced and biased in favor of the controlling agencies from the outset.

Modern media permits propagandizing on a level and to an extent never before seen in human culture, and ideologues have rushed to embrace its use for explicitly political purposes (Helfenbein & Shudak, 2009). Thus, modern political discourse is shaped not by the populace, but instead by those who own and thereby control the mass media, and the interests of such controlling entities overwhelmingly trend toward the authoritarian. As a consequence of this, in turn, the very terminology and vernacular surrounding democracy and the public good has been altered to conform more closely to specific regressive values; "democracy" is now defined as "choosing leaders" (who must then be dutifully followed) and "free markets" are the antidote to the undesirable "socialism" which would annihilate them. And yet, "democracy" is much more than picking and then obeying pre-approved candidates in a representative republic, whereas "free markets" are little more than an unfettered embrace of the latent merciless, anti-social, greed-based hostilities of a competitive economy *in lieu* of showing a little compassion and charity toward one's fellow human beings. Such thought-provoking perspectives are not taught

in present-day schools, by design, and thus real power is kept out of the hands of the governed (Popkewitz, 1996).

Whereas the primary and secondary school systems are perhaps more beholden to government oversight and control, higher education will need to take the lead on the rectification of this problem (Popkewitz, 1996). Higher education is (at least for now) subject to less legislative and regulatory restriction (*cf.* "academic freedom," *etc.*), as well as serving as the primary clearinghouse for the certification of education professionals — teachers and administrators — for the public school system. It therefore falls especially heavily upon higher education academics to confront and redress the need for corrective policymaking at all levels of the educational system, not only by research and study of the problem, but also by first and simply raising awareness, both inside and outside of the profession, of the very existence of the problem in the first place (Helfenbein & Shudak, 2009).

That the governed – either the general citizenry or the professoriate, as the case may be – acquiesce to this is a direct consequence of their systematic dehumanization in the authoritarian-controlled educational system — they implicitly accept a regressive worldview as natural, and even desirable, and subsequently fail to actualize their own potential excellences. From the perspective of educators and the philosophy of education, this may be viewed as an unethical practice on the part of both the society and the individual.

Dehumanizing Consequences

Excessive reliance upon standardized outcome assessment methods may be — in the vernacular of systems analysis and best practices — "considered harmful" (Wallace & Graves, 1995). One of the most insidious consequences of standardized assessments is the normalization of failure. Standardized assessments are inevitably calibrated — or "curved," to use common

terminology belying the Gaussian origins of the practice — not only to generate a "normal" level of performance around which the majority of "non-deviating" students are clustered in the results of the analysis, but also to insure a pre-determined percentage of students have performance metrics low-enough to be considered "failures" at whatever the assessment is designed (or less charitably, in the case of quantities such as IQ, "purported") to measure.

This practice then engenders an "acceptable" number of resultant failures, and such are perceived as an inevitable and unavoidable natural consequence of whatever humanistic process, such as education, has been subject to the assessment (Wallace & Graves, 1995). It should not be overlooked that "free market" ideology itself furthers and perpetuates this inescapable consequence of competition: The existence of "winners" is predicated upon the existence of "losers" to provide context, yet this process, though arguably natural (*cf.* "The Law of the Jungle" and so on), is neither necessary for nor appropriate to the existence of an allegedly-civilized social species such as mankind.

In education *per se*, there are two distinct, but related problems with this (Thayer-Bacon, 2008). Although the accommodation of outliers performing significantly above "average" is often haphazard within educational systems, it is not as urgently needed, as such individuals tend to self-actualize on their own initiative regardless of environmental aids or hindrances, building self-governance models out of whatever is available. For educators, the primary challenge facing these "gifted" individuals is the outward socialization, into a democratic practice, of the principles of self-governance in worldview. Without adequate guidance and context, "gifted" individuals can fall under the influence of competitive, regressive values systems, and never aspire beyond self-centered merely-libertarian perspectives into a full actualization of the

principle of "liberty" as "liberty for all" within the context of a mutually-respectful social structure.

More problematically, the accommodation of outliers performing significantly below "average" is practically non-existent within the mechanisms of governance in contemporary society. Recently, the mass media have been beset with widespread editorializing calls for policy reform in both the social and educational systems because those systems have been producing individuals prone to expressing their personal frustrations and failures through fatally-destructive acts of public violence against innocent fellow members of their communities; be they family members, friends, or strangers, such victims represent targets of opportunity for the dehumanized "loser" since the real instruments of oppression are too abstract and too powerful for the lone (*i.e.*, improperly socialized as a result of regressive social practices) individual to oppose directly. Again, the debate is being shaped by the mass media to serve the interests of the controlling governance; the editorializing is often long on blame and short on plausible, practical remedies, largely from the deliberate avoidance of engaging with the real issue: the systematic depersonalization and dehumanization that has become the accepted *status quo* of modern society. Until the roots of the problem are brought to light — in and through educational reform — social reform, no matter how strongly needed, will simply not be possible (Giroux, 2007).

Such reform, however, is not without its own challenges (Rose, 1996). The lack of centralized mechanisms of control make governance of less-authoritarian models — be they in education or society — much more complex undertakings, and ones which require constant adjustment and self-assessment. In practical terms, this makes them less efficient to administrate and operate, but it may be noted prosaically that this is the price of progress, and that the flourishing and added value such practices bring to both the resulting societies and the

individuals whose lives and works make up those societies combine to make the undertaking worthwhile.

Summary and Perspective

From my training as a philosopher and my experience as an educator I have evolved an admittedly-progressive conceptual framework over the years to contextualize the relationship between education and society. Analysis of this conceptual framework itself then gives rise to a practical framework informing my career goals and educational philosophy.

Within my conceptual framework, I have evolved an epistemological framework best characterized by the works of Michel Foucault (1972, 1994) and Paulo Freire (2006). These two post-Marxist theoreticians are outspoken critics of repressive social orders, and Freire in particular is a pioneer of *critical pedagogy*, a complex educational movement intended to promote critical thinking and social justice. Critical pedagogy invites students to examine notions such as *social space* and the resulting *genesis of (social) classes* (Bourdieu, 1977), and how social power dynamics shape and determine society (Arendt, 1998).

Philosopher John Searle (1990) has suggested, especially in critique of Harold Bloom (and perhaps of the late Howard Zinn as well), that the purpose of critical pedagogy is ultimately to create political radicals. From my perspective since the turn of the millennium, given the current social and educational climate as routinely reported in the mass media and trade press, the intended pejorative quality of this admonition is less than persuasive. The history of human civilization finds political radicals at the fulcrum of many of its turning points, and this is as often as not a valuable and beneficial thing – as it often leads to *progress*. Arguments against progress — no matter what their provenance — are inevitably flawed, being self-serving and/or oppressive (Nietzsche, 1887).

A central tenet of critical pedagogy is the properly-liberating quality of education and the consequent/concomitant benefits to the student and society, hence the traditional appellation "liberal education." The dimensions and degrees of this liberation are defined but not delimited by the accompanying theoretical framework I have erected for myself upon the aforementioned works of John Dewey (1900, 1916) and Henry A. Giroux (1983, 1990, 1997, 2003, 2005, 2007). Yet — as critical pedagogy itself will attest — theory is of only academic use without an accompanying praxis, and therefore in recent years I have become aware of the need to evolve a practical framework for the advancement of education along liberating, rather than oppressing, conceptual lines.

I have identified qualitative assessment methodologies as being acutely wanting and neglected in this regard, as they, by nature, appear to me to tend to empower students and educators rather than institutions of top-down control. My research into qualitative assessment, initially begun out of merely professional interest in improving my own teaching, has grown to become the cornerstone of my entire pedagogical and philosophical outlook toward education.

The occasion of writing my doctoral dissertation provides an excellent opportunity to explore remediative technologies (to borrow vernacular from Dewey) to facilitate the process of progressive educational reform — specifically in line with the problematic trends observed in higher education at present. Both in the literature and out in the field, I have begun to more-frequently encounter lamentations of the increasingly-dire state of education; more and more it seems that everyone is calling for something to be done, but few workable solutions are being offered, short of those that transfer more control of education out of the hands of academe and the *publis* and into the hands of self-serving institutions of not-directly-accountable economic-social-political power.

Although it would be presumptive to expect this investigation to yield a wondrous panacea for the problem, it is my intention that by addressing what I perceive to be a serious omission — either from deliberate, political design or intellectual fatigue and inertia — in the understanding of good educational practice, I may at least gain personal insight into a promising path for educational reform. Whereas it is explicitly not my intention to generate a harsh polemic in this dissertation, it is my explicit intention to construct recommendations and suggest a constructive course of action. If, as William Butler Yates famously stated, "Education is not filling a bucket, but lighting a fire," I note that such fires may be effectively kindled not only in the minds of students, but also in the minds of their teachers, administrators, and public officials.

Best Practices

In the next chapter, a methodology will be developed to establish the parameters of best practices in assessment reform through case studies and applied systems analysis of qualitative assessment. A special emphasis will be placed on the progressive and ethical considerations of increasing reliance upon qualitative assessment methods and rubrics as technologies to serve educational reform better than a purely quantitative, standardized apparatus can.

CHAPTER 3: METHODOLOGY

General Considerations

This chapter presents and discusses the methodological considerations of case study research both in general and as specifically applied to this research task. As such, it starts from a philosophical perspective, elaborates into practical issues, and concludes with a mixture of the two. Subsequent chapters consist of the evidence and analysis, concluding with an overall discussion of the findings and their implications within the framework set out here below.

During the course of a grounded-theory research project such as this study pursues, the methodology may naturally evolve. As my investigation unfolded, the initial, expected emphasis on case study and interview methodology originally proposed below necessarily shifted to an investigation based primarily on document analysis and, due to the poor condition in which some of the documents were found, I ended up developing a more informal coding scheme than the one I had originally envisioned using. These specific changes are further detailed in an addendum to this chapter.

Philosophical Assumptions

The philosophical school of *positivism* informs both educational and grounded-theory, case-study research (Gall, Gall, & Borg, 2006). Positivism centers on the essential *knowability* of reality and has been highly instrumental in shaping the modern scientific worldview; it carries with it its own methodological strengths and weaknesses, however. Based upon reasoned epistemological conclusions about the possibility and obtainability of empirical knowledge,

positivism, if relied upon uncritically, can lead to unwarranted assertions of *certainty* (Popper, 1959). This in itself forms an underlying theme of the preceding chapters: data may be misconstrued as facts, and quantified constructs may be reified out-of-context and thus harmfully. All investigations pursuing a grounded theory strategy — wherein theoretical constructs are typically emergent from the researcher's consideration of the data — would therefore likewise be well advised to adhere to Occam's Razor (also popularly known since Occam's day as The Principle of Parsimony, from the original Latin *lex parsimoniae*) and not multiply entities beyond necessity when theorizing or interpreting, and the present research must be no exception to this — arguably an original best — practice.

Phenomenology in internal context. The discrete *phenomenon* taken under study herein is *effective qualitative assessment*, where "effective" is used here to mean some forms of objective validity and reliability result from the particular practice. Furthermore, within this type of grounded-theory research, a *case* is a discrete and particular instance of the phenomena under examination (Gall et al., 2006): here, a program, instrument, or practice of student learning assessment (a) at an accredited postsecondary educational institution; (b) that relies fundamentally upon a qualitative perspective rather than a quantitative one; and (c) has itself been the subject of expert evaluation prior to this study.

The *units of analysis* for the study are emergent from the particular cases, and naturally evolve into a *coding scheme* as a dynamic part of the investigation. Similarly, the *focus* of the study is also emergent from the basic notion of "outcome."

One of the primary research goals in this study is the evolution of a coding scheme that will facilitate a subsequent (later-study) *factor analysis* of the emergent factors characterizing qualitative instruments such as rubrics. It is the researcher's intent that this study may form a

pilot from which *quantitative* interpretations of relevant *qualia* may in turn be drawn, to lend an appreciable aspect of measurability to the employed qualitative instrumentation.

Phenomenology in external context. As the nature of the task is *interpretive*, the role of the researcher in qualitative research is inextricably linked to the research itself, given the inherently *subjective* nature of the mode of inquiry (Patton, 2002). Therefore, within this type of study especially, explicit attention must be paid to discriminate the *etic* perspective of the researcher from the *emic* perspectives of the participants and subjects. Given that the researcher may be presumed to share a command background and interest with many of those whose work forms the object of the study, the differences in these perspectives may be subtle, and perhaps even irrelevant to the basic details of the investigation. In the summary phase of this research, however, the distinction is non-trivial, and is duly noted as necessary.

In the interests of necessary full disclosure, I again acknowledge a proprietary concern in matters of educational practice as they exist within the larger socio-cultural context in which education and educational systems function. It is my intention that by drawing a clear distinction between the political and social concerns motivating the research interest on a personal and philosophical level and the actual data that may be obtained, a more scientifically-objective framework may be established for the evaluation of qualitative assessment practices in and of themselves. As indicated in the previous chapter, I make no secret of the intended significance and relevance of this study, but — as per Occam — I must restrict interpretation of the findings to only what is demonstrably in evidence from the data and its analysis; researcher self-awareness is necessary to temper any excessive subjectivity to the interpretation of the findings and their meanings. This will be further addressed in the analysis and conclusions in the final chapter.

Role of the Researcher

Peshkin (2000) characterizes the issue succinctly, "The essence of case study design is interpretation" (p. 7). From this, four parameters of interpretation may be identified within the case study format: (a) where the researcher looks; (b) what data is collected; (c) what data is analyzed; and (d) what meaning may be gleaned from the analysis.

Where the researcher looks in case study research is usually informed by convenience and snowball sampling, and the present research is no exception. Methods such as rubrics and writing programs were chosen as initial points of investigation precisely because I was already familiar with the basics of rubrics as an instrument and written evaluations as a process, and had ready access to background information on recent developments in the theories and practices of each as well as a basic conceptual familiarity from which to outline a preliminary coding scheme. Further cases then develop as an outgrowth of these lines of inquiry, as well as any new lines of inquiry that present themselves to the opportunistic, seeking researcher as the investigation develops. For example, my initial interest in rubrics, sparked from casual conversation with an instructor colleague, led me to identify an educator with experience in the successful application of rubrics to writing. That individual also pointed me toward recent innovations in peer-review systems of writing assessment, where I identified two other educators with experience and contacts in that technology. All of these individual educators were then approached about the possibility of being research subjects in a dissertation project, and they not only expressed interest, but also subsequently suggested other individuals and programs for the researcher to investigate. Combined with the lack of specific literature on the matter, this word-of-mouth process quickly convinced me of both the feasibility of, and the acute need for, this study.

It should be noted at this point that in educational research of this type, assessment of *processes* is at least as important as assessment of *outcomes* (Whitson, 1998); the two may be considered inexorably intertwined, and in this study are appropriately evaluated as such.

In most forms of qualitative social sciences research the scope and extent of the data that is collected typically exceeds the scope and extent of the data that is actually analyzed; this can be problematic if the researcher eliminates from consideration — either through accident or design — data that may hold relevance to the inquiry (Creswell, 2009). For this reason, case studies such as this one require review and evaluation from peers and experts at all stages of data handling, to ensure that relevant data — or relationships within the data — are not overlooked by even the most self-aware of researchers. In addition, such external quality control is necessary to establish and maintain a level of *trustworthiness* in the researcher, lest objectivity become compromised from the collapse of etic-emic distance.

Finally, the ultimate interpretation of the data is subject to immeasurable bias from the researcher's own etic perspective, and again, ongoing appraisal from peers and experts can identify and ameliorate this factor beyond what even the most self-conscious researcher may achieve acting only independently, even in light of *full disclosure*.

Overall, I am informed from a philosophical background that looks to Immanuel Kant and Aristotle for ethics, and to Plato, Kant, and John Dewey for epistemology. I am particularly influenced by the aspect of Kant's Categorical Imperative that morally requires each person to ever be considered as an *end*, and never merely as a *means to an end*. This is combined with Aristotelian notions of an internal *entelechy* (loosely, *intellect*) in all things driving and directing each of them toward the virtuous expression of their particular *telos* (loosely, *meaningful purpose*).

I recognize that these elements are inherently present in all students, engendering a Kantian socio-ethical obligation among civilized peoples to not dehumanize students by reducing them to mere cold, alienated statistics for some practical purpose, nor to unethically obstruct the Aristotelian actualization of the unique potentials within the student. The onus is upon educators to facilitate the excellence of each and every student; as discussed above, quantitative assessments contribute little to this underlying moral purpose of education, and may even obstruct it when misused (either by accident or design).

As regards epistemology, I hold a position that acknowledges the universal nature of the objects of knowledge — consistent with Platonism — but, transcending Platonism in the tradition of Dewey, recognizes that knowledge must be individually constructed by each learner, rather than merely recognized as some instance of a divine absolute (a Platonic *aeon* — traditionally translated as Form — as it were). To this process I also add the Kantian perspective that although all knowledge bears this subjective character, the underlying framework upon which knowledge is constructed has a universal consistency (Kant's *categories of the understanding*) that renders some measure of coherence to its resultant understandings — this leads to the aforementioned *positivism* underlying the scientific method, for instance. (Positivism has delimitations of its own, as any philosopher of science will attest, but they are beyond the scope of this present work, and positivism remains a cornerstone of educational research regardless.)

Similarly, this positivist, constructivist perspective has convinced me that, although the process may be unclear at the outset, the research questions pursued in this study are, in both theory and practice, answerable. Furthermore, the ethical concerns raised in my mind regarding

the proliferation of quantitative methods as the preferred basis of educational policymaking effectively require me to address them publicly.

Strategies and Procedures for Data Collection and Analysis

Presentation of the data in the case studies is guided by three processes within the data collection and assimilation itself: description, explanation, and evaluation (Gall et al., 2006).

Description. The bulk of the case studies consists of *thick descriptions* of the phenomenon (or phenomena) (Gall et al., 2006; Patton, 2002). Herein, it is primarily comprised of a detailed exegesis of the history and systems involved in a case, and is typically fairly extensive. From this observation and critical evaluation, salient *constructs* are inferentially emergent to the researcher; these constructs are then conceptualized and grouped according to their coding — again by the researcher — into *themes* that define the relevance of the case to the research questions. The data collection for this case study research involves a combination of (typically public) document analysis — in the form of reports, rubrics, policies, and such — and personal communications — in the form of correspondences and interviews as necessary to elaborate on the written record and themselves documented in turn within the written record of this report. In any sound, rigorous qualitative study, all details of data collection must be preserved in such as way to be *auditable* by the current researcher or future researchers.

Publicly-available documents do not require much in the way of consent or approval; they are available in libraries, over the Internet, and sometimes even by personal request from scholarly researchers. Within this study, it is the correspondence and/or interview process that particularly needs IRB scrutiny to protect the educator/assessor subjects. As part of the process of rich data collection, individuals being corresponded with or interviewed may provide, either from prompting or on their own initiative, information from a perspective or in a capacity that is

personal, rather than public-professional — I encountered this phenomenon in casual, collegial conversation with colleagues even before the study formally began, and consider it to be at least as valuable and informative as "on the record" speech from an official capacity.

Although the direct incorporation of such informal data into the study and analysis may be problematic, it certainly may be used to inform the developing, "on the record" lines of inquiry. Therefore, a fairly comprehensive consent form, attached as an appendix to this document, has been provided to guarantee *informed consent* and thereby protect subjects engaging in dialog with the researcher as part of data collection. As there are only a few preset interview questions to be used across the various studies, where consented to and where especially relevant or salient, member-checked edited transcripts of all relevant and/or referenced interviews and correspondences are attached as appendices to this manuscript to further document for the reader the context and background of anything cited in a given case study.

In special cases and where specifically requested, individuals, institutions, and/or programs are anonymized for subjects' protection; pseudonyms are used as necessary and identifying information in the original records is kept only in a "military-grade" encrypted and hash-signed file on the researcher's own computer system (including archives), with the researcher having sole knowledge of the decryption passcode. In general, however, there is a large degree of transparency in the formal data collection of this particular study, as the data tends to come from publicly-identifiable individuals and programs; this promotes the overall credibility and trustworthiness of the research findings.

As to the applied methodology, the key component of the research process herein centers around the coding of the data, after collection, into a meaningful scheme. This process, conducted upon data shortly after it is collected from each source, results in a *triangulation*

dynamic that can direct subsequent data collection. For this study, there are no explicit preconceptions about of what the data will consist (as mentioned, there are only a few specific, pre-written interview questions at the outset, for example), and therefore it is incumbent upon me to "bootstrap" the research questions for each step of the investigation out of the data from the previous step. The research begins from a handful of salient areas of interest, but the path of the investigation is driven by the data as they are encountered. The unpredictability of this process requires me to maintain flexibility when approaching the data, and can cause interpretations of the data to periodically need re-evaluation. This re-evaluation can even extend to the theoretical framework — critical theory and constructivism — driving the research to begin with, and again, as an investigator, I must not be a prisoner of my own ethnography. A good scientist has no unchallengeable axioms. (For the pedantic reader: note that this statement is a *logical* truth, not an *empirical* one.)

This stage of each case presentation favors a *narrative* voice, to present the context in the most compelling and engaging manner — that of a story (Patton, 2002). Again, the personal perspective and unique ethnography of the researcher inevitably informs and shapes this story, but a journalistic tone should be maintained — at least until the concluding chapter. I acknowledge — consistent with Plato's contempt for the arts — the fact that compelling stories are in and of themselves not logical proofs or even rational arguments *per se*, but as there can be no meaning with at least a modicum of engagement — a point to which Plato would grudgingly agree — evoking empathy in the audience is not an inappropriate strategy when attempting to demonstrate the power and utility of qualitative practices.

Explanation. Two types of *patterns* emerge from themes: *relational* patterns and *causal* patterns. As this research is looking to characterize best practices, from the theoretical

perspective, causal patterns are more significant and relevant to the overall goal of generalizability, but from the practical perspective, relational patterns can be more robust and ultimately useful. Systematic relations between constructs and themes are explored to establish plausible effective links from one to the other, and these links are themselves grounded within the context established from description. The *validity* and *reliability* of the research rest largely on the quality of this analysis, and a technique of *constant comparison* within the emergent constructs must be employed to assure relevance and reliability, even if such are only evident *ex post facto* (Glaser & Strauss, 1967), and triangulation processes further strengthen these.

This stage of the case presentation shifts to, and remains in, an *analytic* voice, to drive a more *objective* mode of explanatory understanding (Gall et al., 2006). Value judgments, being themselves judgments of meaning, are more properly reserved for the concluding chapter and its evaluation of the data and analysis.

Evaluation. The finishing process of each case involves the integration of the findings into the goals and conceptual framework of the research question. At the conclusion of the overall study, a second level of evaluative, critical analysis further extends and generalizes this process across all the reported research through a mechanism of *triangulation* (Gall et al., 2006). It is at this point that the researcher returns to narrative voice, the extraction of meaning from the data being framed in terms of the researcher's theoretical perspective and goals. It has often been said that all observations are theory-dependent, and even grounded-theory research is not immune to this fact.

Validating the Findings

From the positivist perspective, then, identification of emergent themes is the primary indicator of validity in this study (Gall et al., 2006; Patton, 2002). Central to the research

question is how *best practices* emerge; the lack of a readily-available definition of such a construct as "best practices" itself is a significant lacuna this research is intended to address.

The specific findings are validated in terms of transferability and generalizability. The former is the more positivist undertaking, as *post facto* analyses can reveal the obstacles and successes within individual cases that delimit them, and thereby establish *construct validity* (Gall et al., 2006). The later is the more-artful and less-scientific task, but is still approachable systematically (Patton, 2002). One of the guiding principles of the explanatory phase of data analysis is to frame constructs within contexts that are themselves well-defined and well-understood, to facilitate transferability of the constructs to other contexts. From this *internal validity*, an *external validity* is then be reasonably extrapolated — in the case of qualitative research, this manifests specifically as the *trustworthiness* and *credibility* of the study. Finally, *reliability* is difficult to establish from a single, essentially pilot, research project of this scope, but within the construct of best practices, a quality of *repeatability* is inherently sought, providing guidelines extracted from the valid constructs and themes to allow the same (or similar) instruments, programs, and/or pedagogies to be implemented in other postsecondary educational contexts, and perhaps beyond.

Ethical Issues

The ethical dimension of this study is threefold. First, confidentiality regarding the individual students whose assessment occurred under the programs investigated in this study should not be difficult to preserve. Any publicly-available reports or program evaluations, for example, appearing as data in a case herein have typically already passed through Institutional Review Board approval processes at their originating institutions, and may already be collectively anonymized for public release.

Second, as discussed above, the identities of individuals providing personal accounts of qualitative assessment practices may sometimes require protecting, depending on particular circumstances and preferences. Many participants in qualitative research welcome the opportunity to tell their stories (as in Patton above), but many others have reasons to conceal their identities, the identities of their institutions, and perhaps even the identities of the phenomena under study (the program names, for example). Allowances must be made for both those willing to be identified and those reluctant to be identified, and each must be referenced appropriately, specific waivers obtained as required, and so forth. It must always be borne in mind that individuals reluctant to openly participate in research may have some of the most relevant and interesting data to offer (Patton, 2002). To reiterate, this is the area most-typically requiring IRB scrutiny to protect the research subjects — the educators and assessors who may or may not be speaking in a public capacity and thereby enjoy the benefits of academic freedom.

Lastly is the issue of the purpose and nature of this research itself, which is driven by critical-theory-based concerns regarding the ethics of current trends in assessment practices and the consequences of such trends upon education and society at large. As this viewpoint has been openly addressed in the early chapters of this report, *full disclosure* by the researcher has been duly established, and shall be maintained, as the etic and emic perspectives are explored.

Summary

To facilitate coding and analysis, I have chosen to use the open-source Text Analysis Markup System and its associated TAMSAnalyzer software.[1] TAMS produces machine-and-human-readable text-based coding and analysis, and also can be used to indirectly annotate

[1] The source code, documentation, and compiled applications are downloadable from http://tamsys.sourceforge.net/

audio-visual media. As my investigation concentrates primarily on document analysis and interviews, this tool allows me to freely cross-reference and analyze any and all significant terms or themes I identify regardless of the source media. The software is capable of tracking numerous emergent themes across large numbers of coded data points, and thereby fosters not only analysis of the data, but meta-analysis of the analysis itself. The TAMSAnalyzer software package also produces output in a format suitable for displaying graphically, as may be useful.

The dynamic process of coding and analysis forms the basis of my investigation, and shapes my emergent theories and research questions. As mentioned in Chapter 1, my goal is twofold: I am seeking to establish some objective, practical guidelines for the successful employment of qualitative assessment techniques (in higher education specifically, but hopefully generalizable beyond that realm) as well as identify specific qualitatively-based assessment practices that by extension preserve and promote the general purposes of liberal education.

To this end, I am seeking to evolve coding schema not only with an eye toward devising specific practical recommendations to facilitate higher efficiency in and wider adoption of qualitative assessment methodologies in the face of comparable quantitative assessment instruments, policies, and programs, but also with an eye toward identifying themes and outcomes that have relevance to the socio-political theories motivating my research from the broader perspective.

It is my intent that this research may serve as the foundation for further, subsequent research to extend these preliminary findings, eventually putting qualitative research on an equal pedagogical footing with the quantitative research that nowadays is so dominant in educational practice.

The Cases

The next chapters present the particular case studies, culminating in the summary chapter which presents the conclusions and recommendations of this report.

Post-Research Addendum

Case study methodology was not as applicable to this investigation as I had originally anticipated it would be. In addition, although my initial interview subjects provided helpful insights into practical matters – especially with regard to rubrics as they are used in practice – my pursuit of further interview subjects willing to address political and administrative issues that qualitative assessment programs face was not fruitful. This was for a variety of reasons, all of which are relevant to the conclusions and recommendations that emerged from my research. In addition, as part of my graduate studies, I unexpectedly obtained a compensated position as a research assistant in the Ole Miss Center for Writing and Rhetoric, and this provided me with some first-hand experience in the application of rubrics to program assessment, further driving the direction of my research and the evolution of my research questions, and this turned out to be more informative to my investigation than I had foreseen.

Case studies in qualitative assessment ended up being difficult to pursue because of the dearth of long-term, successful programs available to serve as subject cases (Miller, 2012). This problem is the result of a combination of factors I address in Chapters 4 and 5, but briefly: 1) historically, qualitative assessment programs are dynamic and often change – necessarily or otherwise – in response to both internal and external factors, and this process of change is inevitably highly institution-specific and not particularly transferable due to its anecdotal nature, and 2) the politics surrounding qualitative assessment programs are generally difficult to navigate, and consequently can frequently be damaging to careers, reputations, and perhaps most importantly, institutional culture.

Even though I had sought and received IRB approval[2] for my basic interview protocols based on the principle that research into the effectiveness of educational programs does not typically pose risks for participants, I did not anticipate that there would be as much potential danger of personal harm to individuals speaking non-anonymously and on the record about what emerged as such a contentious and challenging issue. One successful interviewee in particular was able to provide some highly-relevant and timely insights into the current state of at least one large-scale assessment program currently underway in American higher education, but due to the commercially-proprietary nature of the technology involved, the interviewee and I agreed that I am obligated to keep all of that data off the record and out of print. This restriction, however, did not interfere with my subsequently seeking publically-available literature that further explores the issues that interviewee raised, and then introducing those findings into my research from the document source instead.

Furthermore, the two problematic factors above can make individuals cautious about speaking directly to their experiences with qualitative assessment programs, and on several occasions, would-be interviewees politely deferred during the initial contact and instead directed me to published essays and reports (and in some cases, books) that reflected or at least were consistent with their own experiences as educators and administrators and either directly or indirectly addressed the avenues of my research while allowing my would-be subject(s) to avoid becoming closely involved with my research. Most of these documents were peer-reviewed, but there were also valuable personal insights (not always centrally-relevant, but useful in providing context and background) to be gleaned from the more editorially-slanted monographs. Consequently, I ended up doing significantly more reading of personal screeds than interviewing

[2] Ole Miss IRB approval number 11-120.

of subjects, but I am convinced that the quality of the resulting data is, if anything, higher as a result.

Likewise, when I encountered corroborating experiences in my own professional experience with qualitative assessment at the Ole Miss CWR, the direction of my research was frequently further informed and revised, and I was able to identify new avenues of exploration in my document-based research as well as receive basic confirmation of some of my fundamental findings. All these considerations shaped the final study, as detailed in the next chapters.

CHAPTER 4: STUDIES IN ASSESSMENT PRAXIS AND ESSENTIAL QUALIFYING TECHNOLOGY

"Conceptual simplicity, [with] structural complexity, achieves a greater state of Humanity" (Shirow, 2004).

Overview

To study assessment without wasted effort, it is important to first establish a conceptual framework to address what assessment *is* and how and why assessment is *meaningful*. Once such cognitive and contextual parameters have been set out, considerations of how the practical tasks of assessment are pursued – here, addressing qualitative assessment specifically – may then be undertaken more fruitfully. Accordingly, this chapter looks first at the theoretical structure of assessment *per se*, and then studies in depth the *rubric* as a highly-adaptable tool with an effectively-unbounded range of applications appropriate to qualitative assessment purposes and situations. From this analysis of the rubric, a general *phenomenological ontology* – a conscious understanding of reality derived from the application of reason to experience – of the assessment process may be conceptualized, both internally and externally to any given application.

The Parameters of Assessment

"No truth explains itself" (Holt, 2012, p. 132).

A fundamental methodological principle of epistemology – dating from Socrates and earlier – is that a necessary precondition for any rational inquiry is that the subject of the inquiry have, or at least be temporarily assigned, an identifying cognitive content at the outset of the

investigation – *i.e.,* one must have at least have a discrete, working definition of *what* one is investigating before one begins the investigation, in order to meaningfully delimit the investigatory process. This definition can potentially change and evolve as part of a feedback loop that may or may not develop during the investigation, but with the (to date) singular exception of problematic Hegelian-style dialectical, completely-emergent phenomenology (which likely lies well outside the scope of something pragmatic such as the present study), the questions of *essence* and *existence* cannot be considered in any useful depth without determining at least some axiomatic notions of one or the other to serve as a starting point for the inquiry. This is a known, and perhaps unavoidable, limitation of the faculty of human reason – whether that reasoning is inductive or deductive (or even abductive, in the logical-inference sense of the term) – and therefore, in accordance with established practices in critical thinking, I must begin by seeking a clear conceptualization of what assessment *is* before I can reasonably proceed further to considerations of how a particular method of assessment may be performed *well*.

In addition, my investigations have led me to conclude that many, if not most, of those in academia who have engaged successfully with the issues of learning/outcome/program assessment have identified some of the key pieces involved (and this often coming only after hard-fought struggles), yet no one I have encountered seems to have successfully contextualized the whole endeavor of qualitative assessment into a coherent, integrated conceptualization.

One reason for this lies in the principle that systems theory describes as *incompleteness*, and the problems of *undefinability* that result therefrom. The incompleteness theorems first published by Kurt Gödel in 1931 that show the limitations of syntax within first-order arithmetic – in part by demonstrating undefinability – have been generalized to formal sematic systems by later thinkers such as John von Neumann and Alfred Tarski. In brief, in any formal system of

symbolic representation of cognitive content (*i.e.*, a language of some sort or other), there are propositions which are not provable within that system, and this is in part because the meaning of the symbols used is always context-dependent upon the system containing them. Whereas this has been proven in purely-mathematical systems, is has also been demonstrated that as a corollary of said proof, the semantical concept of *truth* cannot be encoded entirely within *any* given language or formal system, but instead requires a *metalanguage* capable of applying its own transcendent predicates and positions to the objects of the first-order language (Tarski, 1983).

As a consequence of this second-order principle, purely *top-down* or *bottom-up* epistemologies are each inherently limited. In the case of top-down models, *essence* precedes *existence* – in other words, the model has a built-in prejudice in terms of what may be known – whereas in bottom-up models, *existence* precedes *essence* – in other words, the epistemology can only address what is already present in the phenomena. What is called for then, is a blended, *holistic* approach to phenomenological ontology that combines the strengths of each traditional epistemological strategy to compensate for the other's weakness, allowing a phenomenological ontology to emerge.

To date, one of the most enduring, landmark examinations of the parameters of assessment is found in the work of Patrick T. Terenzini (1989). In addition to examining the basic logistical issues faced by postsecondary institutions seeking to improve assessment practices, Terenzini helpfully lays out an analytical framework for studying student outcomes, and a brief summary of his taxonomy is worth including here.

As assessing educators, we must begin by addressing three general questions, plus a set of particulars within each of those questions. First is the question of *purpose*: why the assessment

is being undertaken (Terenzini, 1989). Although this may seem a pedantic issue, it is genuinely informative as responses will generally fall into one of only two categories: assessments of teaching and learning, which may be construed as *formative* matters, and assessments of accountability, which may be construed as *summative* matters. Although it may be argued that subsequent advancement of assessment models in the intervening years since the original publication of Terenzini's work has in some sense collapsed this distinction to the point that summative assessments such as *program assessments* may now be considered simply a direct extension of *learning assessments* (the other metric for program assessment being *cost efficiency* and/or *cost-benefit analysis*) the formative/summative distinction remains useful for purposes of understanding and justifying assessment efforts in terms of their ultimate purpose and utility; formative assessments drive program modification and improvement, whereas summative assessments inform judgments of program worth or value.

Second is the question of *level*: who is to be assessed (Terenzini, 1989). Here, the distinction is drawn between individual students and aggregate groups of students. Traditionally, assessing individual student outcomes is handled by instructors assigning grades, but programs are more likely to be interested in aggregate student performance against some standard for summative purposes. Terenzini notes that there are numerous grouping criteria that may be relevant to program interests, be they either organizational (course, program, department, college/school, campus, system) or demographic (gender, ethnicity, class year, major, residence, *etc.*).

Third is the question of *object*: what is to be assessed (Terenzini, 1989). This is the most complex and challenging question of the three, and Terenzini points to a broad four-fold typology provided by Ewell (1984) as being both simple and comprehensive: 1) *knowledge* (both

breadth and depth), 2) *skills* (including basic, higher-order, and career-related), 3) *attitudes and values*, and 4) *behavior* (both during and after college) (Terenzini, 1989).

Together, these three dimensions from Terenzini and eight (sub-)criteria from Ewell can encompass the entire scope of postsecondary assessment as it has been and might foreseeably be practiced (Terenzini, 1989). Of greater interest to the present inquiry, however, are the various problems that even well-designed assessment programs may run afoul of. In examining the literature of case studies and analyses of assessment, I have developed a general taxonomy of the problems assessment programs face, and an overview is presented here to serve as a basis for my further analysis and exegesis.

In summary, and as to be detailed subsequently, assessment problems may be categorized into four basic types. The first and foremost of these are *conceptual* problems best addressed by Terenzini's analytical framework: all stakeholders need to share a clear, common, and distinctly-articulated set of ideas and expectations regarding any program of assessment that is undertaken, and although different stakeholders may have particular vested interests – to which varying degrees of significance and importance may be appropriately assigned – and whereas the development of some of the specifics of this conceptualization must be an ongoing and concurrent developmental part of the assessment process, having the conceptualization of the assessment firmly grounded from the outset is essential to the assessment's overall meaningful success.

The second area of problems concerns *measurement*. This is a dual issue; at the core, *observer effects* – originally from quantum physics, but nowadays extended to the social sciences especially – can be difficult to counter inasmuch as modern scientific methodologies acknowledge that even when taking into account other internal and external influences, anything

that is measured may be observed to change over time in response to being measured (López, 2002). More problematically however, measurements of student learning may be either *direct* or *indirect*. Whereas conventional instructor-administered subject-based examinations and tests remain an essential part of teaching pedagogy, for broader assessment purposes direct measures of learning are typically *performance task-based* and are characterized as being "authentic" in the sense that they require students to solve realistic problems that are unstructured and have no explicit "right" answers (López, 2002, p. 362). These direct measures are generally applied at the individual level, at least in their initial data collection. Conversely, indirect measures focus on the "perceived" extent or value of learning experiences and typically include instruments such as surveys, cohort studies, exit interviews, and trend analyses, and their application is concentrated more at the group level. Although the direct measurement scheme is obviously the more immediately pragmatic of the two, both direct and indirect assessment measures that go beyond the conventional written-answers-to-explicit-questions "test" (as such) are crucial to bringing meaning to the assessment process by showing educators not only *that* students have learned, but more essentially *what it is* that students have learned. Making this distinction is imperative when discussing the purpose and goals of student testing at all levels of education.

The third problem area for learning assessment programs is the *organizational*, and perhaps unsurprisingly this consideration also applies at more than one level: as with any initiative successfully implemented within a organizational environment, an assessment program must be internally organized and aligned with its goals and purpose, while simultaneously being externally organized to function as an organic part of the institution it serves (Rossman & El-Khawas, 1987). Each of these two organizational aspects is equally crucial to the success of the program.

The fourth and final problem area I have identified is, perhaps as might be expected, the *political*. The institutional culture typical of higher education embraces a unique combination of anything-but-unique administrative and social relationships which can make or break any project and the people involved with it, often irrespective of the relative merits of either (Shatzky, 2012). This is perhaps the most complex and subtle of the four problem areas, but there are a myriad of recognized ways to engage it, if the will can be found.

Instituting "Best Practices"

The notion of *best practices* might be uncharitably characterized as being "…indistinguishable from mindless mimicry, the very opposite of academic discovery and insight" (Cooper, 2012, para. 8). Though such a dismissive judgment is perhaps overly harsh, it also may be indicative of the gulf that has developed in recent years between what has now become the default perspective of administrators and that which has traditionally been held by faculty in regard to what is valuable in higher education. Both factions may rightfully be viewed as being *results-oriented*, and yet they may find themselves at odds over what results are to be desired and prioritized within the institution. In the literature and in interviews, I have found that best practices are often recommended solely on the basis of their ability to reliably produce the desired outcomes, but these recommendations are often free of contextual or theoretical groundings that do not rely upon some form of circular reasoning backwards from those outcomes to begin with. The data is taken, and outcomes are defined and delimited by what may be gleaned from that data, but rather than forming a dynamic, evolving feedback system, the data may begin to drive the parameters of assessment by themselves. Ethical and pragmatic considerations demand that this reflexive mechanism be improved upon in order for the assessment to be authentic and meaningful instead of merely self-serving.

Ultimately, any recommended best practices will naturally share certain essential properties derived from practical experience that identify those practices as being effectively "best" (and oftentimes this may simply be the result of trial-and-error testing in an environment of institutional Darwinism), the foremost among these being *repeatability* and *transferability* (hence the "mimicry" pejorative above). The principles of validity and reliability, as applied to the analysis of assessments and their instrumentation, can support this kind of portable utility and thereby establish foundational qualities for effective and valuable assessments to have in common, even if the assessment programs themselves remain highly institution-specific (Terenzini, 1989). I shall first address the foundational issues of how to establish qualitative assessment best practices in the context of "bottom-up" technologies, and then address the contextual and environmental challenges that assessment programs face using "top-down" methodologies; in this way, each of the two aspects can compensate for the limitations of the other, and as a result a coherent, integrated, and grounded theory can be realized.

Ewell (1984) points out that the "appropriate" outcomes of higher education were once "relatively few and well agreed upon" (p. 18)[3] inasmuch as they consisted of a straightforward

[3] The only available extant copy of the document appears to be a PDF composed of scans of photocopies; the pages of this PDF have been numbered sequentially – often with numerals superimposed on the images – to include other parts of the document history and context, such as copyright and printing notices, database and filing records, and so on. This compiled document apparently now forms the only surviving version of the source text to be found. In many cases – in this source and ones in similar condition – the original page numbers also are sometimes obscured or missing in places, making the version prepared for archiving the most-reliable version. Thus, I cite the revised page number here in lieu of the original.

combination of character building through a traditional liberal education with intellect building through scholarship in a chosen academic discipline. In recent decades, however, this conceptualization has changed and evolved in complexity to the point that new dimensions of educational outcomes must now be invoked. Such dimensions may be understood as threefold (and Ewell's order is telling, p. 21): institutional objectives, student educational goals, and the needs of society and other third-party stakeholders. It should not be overlooked that Ewell's notion of appropriateness is closely tied to a principle of *accountability* that was relatively innocuous and virtuous in the early 1980s, but which has taken on a different political character since the turn of the millennium – a character that is now focused on exerting external control.

Ewell (1984) does not fall into what I would characterize as the "trap" of accountability however, recognizing that there is a crucial distinction between the *empirical changes* produced in the student and the *values* placed upon those changes both inside and outside of the academy. As educators, the effective delivery of *learning* to our students is and should be always our overriding concern. As discussed in previous chapters, within the modern political climate outside academe the primacy of this goal is increasingly subject to alteration and repurposing in order to serve special interests at the expense of the public interest. As further analyzed below, the inferences that may be drawn about student learning shape and are shaped by the interests that are being served in actual practice. Although there will and should always be pressure for external conformity in postsecondary outcomes, practical purposes drive the need for explicit, institution-specific goals to be identified and acted toward instead of surrendering that function to calls for and from external conformity.

This distinction then invites the question of what, precisely, a "learning outcome" is or can be (Ewell, 1984, p. 16). The subtleties involved are non-trivial: on the one hand, the change

characterized as "learning" can be construed as a relative increase in a student's knowledge or abilities from their initial state, but on the other hand, the change that carries the label "learning" can be measured against an absolute scale of proficiency independent of the student's initial aptitude level. Each of these metrics may be informative in its own right, but their respective utility depends on the specific analytical and assessment purpose to which each may be applied.

In addition to the issue of establishing the baseline reference against which the learning (or progress) will be construed and measured, assessments (learning or otherwise) may be either *norm-referenced* or *criterion-referenced*. The former refers to comparisons of performance against that of peers (either internal or external to the context), whereas the latter refers to performance as measured against objective standards (which may also be either internal or external to the context). Although elements of both may be relevant to an assessment undertaking, referencing criteria rather than norms is more foundational to the process of phenomenological ontology that in turn grounds the assessment conceptually within the discipline and the institution. Addressing primarily issues of standardized (and hence, more-quantified) testing, Harris (1986) tells us:

> You can compare your students to other students nationally on standardized tests without having definite educational goals, stated expectancies, or outcomes. But without such goals, you can't be sure the tests reflect your curriculum. You and your colleagues may also be interested in how your students change in terms of beliefs, interests, attitudes, values, and behaviors. There are various commercially available inventories to reflect these things. Yet again, without relatively clear student development goals, you won't know how to select the inventories that fit your institution. (p. 22)[4]

[4] Note 3 applies to this document as well.

This necessary contextualization of conventional test-based assessments extends even more crucially to qualitative assessments, especially those developed from within the institution itself.

Although Ewell (1984) asserts that a lack of administration-driven incentives for the faculty is the major obstacle to effective assessment reform, he provides several short case studies of successful programs (and the other effects on their respective institutions) in the second chapter of his report. As his report was subsequently used as a basis for recommendations made by the United States Department of Education to American colleges and universities, the significance and influence of his findings has been considerable.

Most notably, there are three key takeaways whose relevance has not diminished in the intervening decades. First and foremost, Ewell (1984) finds that assessment programs do not necessarily require high costs, specifically a "massive infusion of external resources;" instead, Ewell contends that much of the cost of reforming the assessment process may be borne by realigning existing internal assessment activities (p. 17). I have a few reservations about this assertion, however. Critical thinking invites us to always consider the agenda behind and the intended audience for an argument, and the reader cannot completely divorce Ewell's conclusion from the potentially propagandistic purposes that commissioned his research initially and to which it may have been put. If an implicit goal of Ewell's report is to introduce a call for assessment reform and encourage its embrace, addressing the easiest and obvious objection ("But it's too expensive!") directly makes good rhetorical sense. This in itself, unfortunately, is insufficient to fully substantiate the truth of the claim that reformed, proper assessment is not particularly expensive and merely requires some re-allocation of existing resources. Indeed, the contemporary trade press and innumerable educator weblogs frequently devote a large potion of their attention to the significant resource costs qualitative assessment programs incur nowadays

(see López, 2002, in particular). Writing in 1984, perhaps Ewell is being overly-optimistic and not disingenuous in anticipating that the costs of assessment could be expected to remain constant and modest even in light of unforeseeable changes in technology and (external) stakeholder expectations that drive corresponding change in the whats, whys, and wherefores of learning outcomes.

Secondly amongst Ewell's takeaways is his point that "effective" efforts must be both "institution-specific and participatory in character" (p. 17). This point is well-supported by other research and even other management contexts (as detailed below), but again, the critical thinker might well take note of the following quote:

> Neither government nor the research community can hope to impose solutions – no matter how well informed – if faculty and administrators have not first internalized the logic of these solutions through their own evaluations and experiences. (pp. 17-18)

Although it appears upfront enough, this statement is highly problematical as it reinforces what has now been shown to be a contentious assumption that "top-down" implementation is desirable, or even feasible, as the ultimate driving factor in assessment programs. To a close reading of the subsequent chapters, it emerges that Ewell clearly acknowledges the necessary "bottom-up" nature of the process, but still his choice of the verb "impose" remains. Thus, Gallagher (2011) identifies something from policy literature bluntly labeled "the implementation problem." This "problem" results from the practical fact that "Policy directives… do not execute themselves" but instead must be implemented by those farther down – often much farther down – in the operational/managerial hierarchy from those who issue the directives and implement them by "remote control" (Gallagher, p. 463). Darling-Hammond (1997) notes that such top-down management tactics are especially prone to failure in an environment of academic freedom

due to processes of *interpretation*, *redefinition*, and even outright *subversion* influencing the implementation. In turn, Gallagher explains, neoliberal education reformers at the top of administrative hierarchies respond to the "problem" by attempting to tighten administrative control with forced programs of scripted instruction, packaged curricula, and standardized testing (p. 463). From Gallagher's perspective, all these control mechanisms become in practice *obstacles* to the processes of education and assessment, processes which are in the final analysis based fundamentally upon teacher-student *relationships* and not upon instructional technologies. In addition, Ewell (1984) tells us

> Many educators argue that the very nature of the higher-education enterprise effectively precludes improvement through increased external accountability. (p. 11)

Although these two factors from Ewell and Gallagher may be convincing to some, the "facts on the ground" (to borrow a popular contemporary idiom) in a higher education system that is increasingly beholden to neoliberal ideologies are best characterized by Ewell (1984) himself:

> What we are now increasingly being asked to demonstrate is nothing more than that for which we in the past have had the hubris to claim credit. (p. 12)

And this perfectly encapsulates the crux of the matter.

Ewell (1984) goes on to describe external standards applied to the academy as being "alien" and "counterproductive" as they are antithetical to the very ideas of academic freedom and inquiry – traditionally, academic standards have been maintained by the individual *disciplines* organized within academia (p. 13). I note that the etymology of the word "discipline" is no coincidence here.

Additionally, Ewell (1984) invites us to consider the traditional role of scholarship in not merely building "society's store of knowledge" (p. 13), but in driving civilization forward to new

levels of *self-regarding* and *self-improving*, while simultaneously producing a similar manifest excellence in the individual learners whose efforts build and advance civilization as a whole. As thus envisioned, education is indeed a bottom-up process of individual- and civilization-building.

Tools for the Job

The *rubric* is arguably the instrument of choice to use when pursuing qualitative assessment in any form; Stevens and Levi (2005) describe the rubric with disarming simplicity as "a scoring tool that lays out the specific expectations for an assignment" (p. 3). The advantage of rubrics comes from combining the utility of being applicable to an unbounded range of assessment scenarios with the virtue of establishing objective criteria for assessing – criteria that are themselves subject to review and evaluation. Rubrics are initially devised using a top-down internal methodology in order to delimit and define the outcomes under consideration, but are subsequently effectively applied – and revised as may be necessary – in a bottom-up feedback process directly driven by those who are employing the tool. As Stevens and Levi provide a definitive overview of the process of constructing a rubric, I shall include a brief, annotated summary of their guidelines as framework here.

The process of constructing a new rubric to use in an assessment (or editing/repurposing an existing rubric) first begins with a *task description* (Stevens & Levi, 2005). From the outset then, the instrument aligns itself with performance-task-based assessment rather than fact-recall-based assessment. This focus is appropriate to *qualitative* assessments that explore not merely whether or not students have mastered a body of knowledge, but *to what use* students can put that knowledge mastery. Typically, this part of rubric design is addressed by assigning an appropriate title to the rubric and, space permitting, including an explicit statement describing the task to be assessed. It is important not to omit or gloss this element, as it may be very

informative to both the students and external stakeholders who seek documentation of the assessment process in order to understand the context within which the rubric is, was, or will be applied.

The second part of rubric-building consists in devising a *scale* that reflects various levels of qualitative achievement and that may be used for quantitative analysis (Stevens & Levi, 2005). As this scale is by default an objective one, the associated learning outcomes may be considered standardized. This point is highly significant, as it is essential to rebutting criticisms from the more purely-quantitative assessment camps that qualitative assessment lacks sufficient standardization. The real issue present is that, in the institutional context, this rubric-driven standardizing is driven from the bottom up instead of the top down, and yet in practical terms this does not make any developed standard any less of a standard, merely less of a mechanism for wielding authoritarian control within or over the institution.

Stevens and Levi (2005) make the point that the achievement scale need not be overly-granular; a five-point scale is usually preferable to a ten-point one in terms of the ease and speed with which it may be applied, and my experience suggest that a four-point scale may be better still due to the fact that it forces every rater to actively evaluate each criterion and subsequently score it either above or below the mathematical, midpoint average. Typically, the scoring should be normalized (by design and in practice) in such a way as to generate an average value for each criterion in the middle of the four-point range, but that average only emerges from the data analysis – raters are not allowed to lazily assign it. In this way, the "average" level of performance is more honest, as it is necessarily an artifact that emerges as a consequence of an evaluation yet is not and cannot be an explicit choice by a rater.

Generally, even a three-point rating scale is usable (extending even to Fail/Pass/Honors conventions, for example), but a two-point scale is not (Stevens & Levi, 2005), as scales this short can lack the ability to make meaningful distinctions in performance. If all-or-nothing measures are desired, it is possible to devise component checklists within each criterion of the rubric and then sum them to yield a general level of performance for that criterion, selecting different levels of student competence on each sub-criterion if particular aspects of student performance within the overall criterion vary.

Although Stevens and Levi do not explore it in detail, it should be noted that for alternative purposes performance levels that are defined relative to the baseline (pre-treatment, *i.e.*, pre-learning) proficiency of the individual student(s) are often possible to devise and apply to different outcome considerations. Indeed, one of the advantages of performance-based assessment is that with sufficient data capture (retaining written work, retaining recordings of presentations, *etc.*) it is possible to re-assess historical task performances using completely different rubrics in service of completely different outcome assessment – a possibility largely precluded in conventional, multiple-choice-based standardized testing. For example, at the Ole Miss Center for Writing and Rhetoric we have from time to time in recent years re-assessed our collected archive of student writings in terms of different program outcomes, devising new rubrics as necessary to let us identify emergently-relevant historical trends in different *qualia* that may be found in our growing repository of student writing samples. The subsequent findings of these re-assessments have then informed decisions about pedagogical revision within our writing program.

The third part of rubric-building consists of determining the "dimensions" of the assessment (Stevens & Levi, 2005). These factors should be identified and labeled with *nouns*,

not *qualifiers*, and are chosen by the educator(s) to capture the essential qualities of the performance task as may be formulated in the (learning) outcomes.

It is here that a rubric gains (or not) the basis of its *validity*. The meaningful correspondence – or lack thereof – between the what is assessed and what is learned hinges upon the appropriateness and fitness of the dimensions of the assessment to the sought outcomes of the educational process. I have done some preliminary research into applying *factor analysis* to this part of rubric design to investigate whether or not a rubric's dimensions may be reliably constructed to be functionally *orthogonal* to each other, and the initial results are promising. Yet I only mention this in passing here for completeness, as even the most mutually-orthogonal of assessment criteria/factors may be still be invalid if the design of the rubric is not properly grounded in a sound phenomenological ontology, and so I shall further examine the issue of validity in a moment. Here, I simply note the direct correspondence between the phenomenological ontology that the learning pedagogy is intended to foster in the learner(s) and the optimal structure of the rubric that is devised to assess that learning. This is the crucial aspect of rubric design and use.

The fourth and final part of rubric-building consists of providing the *descriptions* of the various *levels of performance* within each dimension (Stevens & Levi, 2005). Stevens and Levi recommend that the highest level of performance should be explicitly described, with lower levels being subsequently codified in relation to this benchmark as desired, and students (as well as raters) typically report having this expectation set out for them plainly is informative and useful (L. Schrock, personal communication, Spring 2012; S. Flaschka, personal communication, Spring 2012). Rubrics which concentrate only upon specifying the highest level of performance, and rating students in terms of how their performances measure up to this single standard are

typically classified as *scoring guide* rubrics. However, rubrics that describe all the levels of performance – from the highest to the lowest – are more common, and this is because although scoring guide rubrics are simpler to write (Stevens & Levi, 2005), rubrics with more detailed descriptions are easier and faster to use (S. Flaschka, personal communication, Spring 2012).

Time-efficiency is a key factor in all aspects of rubric use. Stevens and Levi (2005) mention that the time invested in devising and writing the rubric is quickly recouped in using it. This is because a well-designed rubric can function as a form of *checklist*, enabling the user to avoid the need to write the same notes over and over from paper to paper (for example). The more detailed the rubric, the faster it is to work through in applied use.

Beyond the emphasis on time-efficiency, Stevens and Levi (2005) also broadly address the general question of *why* rubric use is desirable and valuable, identifying essentially six different elements of a comprehensive rationale. Pursuant to their book's goal of promoting rubric use, they frame their justifications in wholly practical terms without touching much upon the theoretical and conceptual considerations that might also be involved. Such considerations, however, are highly relevant and it is a simple-enough matter to identify and understand them.

Extending the time-efficiency paradigm, Stevens and Levi note that the ability of the grader to rapidly process the assessment consequently facilitates providing *timely feedback* to the student, and this timeliness is viewed as being essential to the learning process (2005, p. 18). Furthermore, as students tend to make the same or similar mistakes on any one assignment, predictable notes may be incorporated into the rubric's dimensions to specifically address these mistakes or shortcomings in the desired outcomes. Stevens and Levi do not explore it in much detail, but this design consideration specifically aligns the instrument – the rubric – with the

assessment of explicit learning outcomes, and the relevance of a rubric's dimensions to its validity arise from this congruence.

In addition to timely feedback, the rubric can also provide *detailed feedback* on performance, as the specifics within the evaluated dimensions can give students categorical explanations for why an assessment was scored the way it was (Stevens & Levi, 2005). Students often report that they struggle in their comprehension not only with course content, but also with instructor expectations, and the more details a rubric includes, the more clearly instructor expectations may be understood (A. Myatt, personal communication, Summer 2012).

Stevens and Levi (2005) suggest that an instructor retain copies – either physical or electronic – of student-submitted originals for the purpose of tracking student performance over time (even using perhaps a different rubric), but personal experience shows that such document retention and preservation practices can also be essential in preventing student challenges to assigned grades based on altered, counterfeit versions of the work offered to one's superiors as "evidence" of grader error. *Caveat lector.*

As an extension of the feedback process, rubrics *encourage critical thinking* by helping students self-analyze (Stevens & Levi, 2005). It is a basic truism in education that self-assessment is a prerequisite for self-improvement, and such self-improvement is often an implicit – if not explicit – learning outcome in its own right.

Equally important, however, is the fact that rubrics *facilitate communication with others* in the instructional process. Stevens and Levi (2005) note that this is particularly valuable in regards to the work of both teaching/graduate assistants who have been assigned to instructors as graders and to the efforts of any tutors who might be involved in the pedagogy. Clear and meticulous communication of outcome expectations to third parties functioning in intermediating

roles between instructors and students is essential to providing an effective education to those students. The crucial nature of this communication may also be extended to new faculty and adjuncts who may lack their more-senior colleagues' long-term familiarity with a given curriculum, as well as among those senior colleagues themselves when there is a pre-established departmental goal of delivering a common curriculum. In practice, this consideration applies primarily to lower-division undergraduate courses more than upper-division or graduate-level ones, and as such, it is even desirable on occasion to have a unified rubric employed department-wide pursuant to curricular consistency in 100-level instruction. In the Ole Miss Center for Writing and Rhetoric, for example, as much of our freshman composition instruction is handled by adjuncts and inexperienced graduate students, a set of core-faculty-devised common grading rubrics is an essential part of our instructional quality control and outcome reliability (or consistency).

In a similar vein, a good rubric can *help refine our teaching skills* (Stevens & Levi, 2005). Institutional and program assessments can analyze scored rubrics to gain insight into the effectiveness of teaching and pedagogy. Whereas some instructors may be put on the defensive at the suggestion that they be judged (partially) on the basis of their students' learning, it should be noted that a *post hoc* examination of scored rubrics can provide more reliable insight into student learning than the traditional, problematical student evaluations of instruction do, having removed the opportunity for students to introduce bias in their responses to instruction (such as might be present, intentionally or not, in replies to survey questions). This is admittedly a more indirect method of assessing instructor effectiveness, yet it also more objective and is therefore more desirable: rubrics provide tangible evidence of student learning (when it occurs). The implications of this for evaluative purposes are clear.

Lastly, rubrics *level the playing field* academically (Stevens & Levi, 2005). Rubrics may be used in any and all disciplines, and provide a mechanism where diverse outcomes in a wide variety of tasks, each variously relative to institutional goals, may be evaluated in a similar manner. Beyond specific course/degree requirements, rubrics can yield insight into cognitive conceptualizations and phenomenology that signify mastery of a subject and are transferable to other disciplines and academic/non-academic applications. This is because rubrics can address cognitive contexts, backgrounds, and biases in ways useful to an administrative understanding of the learner and the learning process. Furthermore, consistent, broadly-applied rubric use can yield a proverbial "paper trail" of a student's learning, and this has relevance at every scale ranging from the individual learner's growth and development as a student up to the institution-wide accreditation process and beyond, documenting the tangible *education* as it occurs.

Within the labor of constructing (or perhaps even merely re-evaluating) a rubric, the process may be divided into roughly four stages (Stevens and Levi, 2005). I use the word "roughly" because experience shows that it is not necessary to be completely pedantic about the task of rubric construction, nor are the stages in practice necessarily as distinct and clear-cut as Stevens and Levi would have them in theory. Nonetheless, the general four-stage method consists of *reflecting*, *listing*, *grouping and labeling*, and finally *application*. Taken in total, this progression forms an integrated phenomenological ontology – one that should reflect in microcosm the larger curricular content.

The foundational exposition of this phenomenological ontology may be found in the questions Stevens and Levi provide to prompt the initial reflection stage. Although the components considered at the outset do not bear any necessary correspondence to the final dimensions the rubric embodies, they can guide and delimit the assessment context as thoroughly

as may be desired (Stevens & Levi, 2005). By analyzing this initial ontology, we may gain insight into the broader value and meaning of the task, the assessment instrument, and the outcome(s) as an educational technology. The questions Stevens and Levi pose to us in the beginning stage of rubric design are as follows:

- Why this assignment?
- What is the history of this assignment?
- What is the place of this assignment within the curriculum?
- What are the skills required for this assignment?
- What is the specific task of the assignment? (And does it have component tasks?)
- What is evidence of accomplishing the task?
- What are your highest expectations of performance?
- What are the failure modes? (Historically or expected?)

It is of course possible and perhaps even desirable to rephrase or otherwise contextualize these basic eight inquiries for particular disciplines or programs, yet upon examination it becomes clear that any conceivable educational or institutional outcome can be subsumed under one or more of these focused questions. The correspondence of the structure of the Stevens and Levi ontology to that put forth by Terenzini reflects an underlying commonality addressed by their respective phenomenologies: by considering, from within a clearly-identified conceptual framework, all the aspects of what the desired outcomes are and how those outcomes may be understood, we as educators can assess any learning process in a reliably systematic manner, and that system will have certain individually-specified, yet generally-universal properties that are shared across inter-disciplinary and inter-institutional environments. This yields the validity of the rubric-as-instrument, as discussed further below.

The subsequent stages of the method all represent refinements and extensions of the basic schema set out in the beginning, foundational stage of reflection. The answers to the reflective prompts then inform the next stage: listing (Stevens & Levi, 2005). The purpose of the listing process is to inferentially identify discrete themes and concepts, from a constructivist model of learning, that an assessment may address. The listing stage is perhaps the most problematic in the process, as it requires the instructor(s) to confront the sometimes-dreaded notion of *learning objectives*. As a consequence, this is often the place in any assessment project where faculty resistance is greatest (Banta & Blaich, 2011; Reed, Levin, & Malandra, 2001; Tagg, 2012), and so I will examine it in more detail below. For now, I simply note that the opportunity to take ownership of the delineation and exegesis of learning objectives represents academic freedom at its most fundamental: the faculty may decide for themselves what shall be taught by clearly specifying to the administration what shall be learned. Good leadership within the institution is required to facilitate the embrace of this perhaps most tedious and pedantic of tasks, yet retaining ultimate authority over learning objectives can form a strong bulwark for supporting a faculty's role in shared governance and maintaining the institution's quality of scholarship.

The listing process is not without its own guidance. In practice, learning objectives flow naturally from contemplation of the eight questions from the previous stage, as well as being informed by curricular/catalog course descriptions and discipline-specific program goals (Stevens & Levi, 2005). Once the learning objectives have been identified and described in a list, they may then be codified in the subsequent stage of grouping and labeling.

Ideally the objectives should be assigned to thematically-similar groups, and the dimension of each group should be as orthogonal as possible to those of all the others; in other words, we want to group like with like and make certain that the resulting groups are distinct and

discrete, with little overlap. This process is more art than science, but may be assisted with formal tools such as the aforementioned *factor analysis* and, most importantly and as indicated above, considerations of the instrument's *validity*. The former element is more *post hoc* to the assessment process and plays a larger role in the revision process a rubric should be periodically subjected to – such revision being central to the latter, less-formal element. Validity is a complex subject in assessment, and there is much that has been written addressing its overall challenges, but for the moment my focus will stay on the technology of rubrics in and of themselves.

Besides grouping the outcomes into an ordered structure, the various levels of accomplishment in performance should be labeled (Stevens & Levi, 2005). In the case of a *holistic scoring guide* rubric, the individual performance criteria for each level should be specified in order to compensate for the potential lack of detail such rubric offers the student in comparison to a more-elaborate *multi-level* rubric (S. Smith, personal communication, Summer 2012). The combination of these two grouping and labeling mechanics then yields the traditional, structured form that gives a rubric its categorical schema. In all cases, designers should bear in mind that the number of groups multiplied by the number of labels yields the total number of elements the rubric will span, and this number should be large (and therefore detailed) enough to make the feedback meaningful but not so large (and therefore complex) as to make the feedback difficult to comprehend.

The final stage of rubric-building (or revision) is the application of the chosen schema to the design and layout of the rubric grid (Stevens & Levi, 2005). Arranging the factors in a grid layout facilitates the ease-of-use and clarity of the rubric by generating a matrix that recapitulates the ontology the rubric is intended to represent, and this correspondence in turn is central to ascertaining the instrument's validity, as below.

Some further detail on the differences between scoring guide rubrics and multi-level rubrics is appropriate here. Of the two, the scoring guide is the simpler design (Stevens & Levi, 2005, p. 39). On the plus side, a scoring guide gives greater flexibility in response, and is better suited to real-time grading of performance tasks. On the minus side, a scoring guide does not provide particularly detailed feedback unless more time is invested later (after the task is completed). For more-comprehensive assessments, a multi-level rubric is preferable (with Stevens and Levi recommending 3 to 5 levels for each criterion, p.79), as such rubrics can combine detail with efficiency. Since devising a multi-level rubric requires specifying a greater level of detail under each criterion, it can be slower to create than a comparable scoring guide, yet the resulting tool is also faster in actual use, since scoring on a multi-level rubric can often be reduced to simply marking the corresponding levels of performance on each criterion in the pre-written matrix. In contrast, scoring guide rubrics allow for greater individualization of the feedback along with greater flexibility in the form that feedback may take. Typically, a scoring guide rubric gives a student a set of "structured notes" as feedback, whereas the multi-level rubric provides a detailed checklist (p.79). Either, or even some hybrid combination of both, may be selected as appropriate to any given task, but I should note that multi-level rubrics more-easily lend themselves to statistical analysis, as they are already arranged in a gridded matrix of criteria and scores within those criteria. It is of course also possible to perform statistical analyses of pure scoring guide rubrics, but some qualitative data coding of the scored rubrics may be required first, requiring more additional time and labor to prepare for analysis than evaluatively-comparable multi-level rubrics do.

It is plausible that scoring guide rubrics are better suited to assessing graduate student and creative work (S. Flaschka, personal communication, Spring 2012; L. Schrock, personal

communication, Spring 2012), but this is debatable (P. Hill, personal communication, Summer 2012; S. Smith, personal communication Summer 2012). In practice, as to the form a rubric might most-usefully take, the deciding factor is typically the time-management of the assessor (A. Myatt, personal communication, Summer 2012). The literature is surprisingly tacit on this distinction and the relative advantages of either option, perhaps because it is a comparatively minor one best accommodated as a particular assessment context may individually require. I include this anecdotal finding from some of my interviewees here simply for completeness, and I note that general experience suggests that the multi-level rubric is better suited for use in situations where the instructor is not the grader – using the more-detailed format, it is theoretically possible to instill some basic *quality control* in an assessment process that relies heavily upon graduate assistants and/or teaching assistants to function as graders by providing a rubric that is detailed and explicit in its outcomes (and expectations) to both those who produce the work and those who grade the work. This then serves as a highly-germane instance of facilitating communication, as mentioned above.

One area of salient interest that Stevens and Levi only barely address is that of *metarubrics* (2005, p. 93). It is highly significant to program assessment that rubrics may be devised to assess the effectiveness of rubric use in learning assessment. In fact, rubrics may be devised to assess any and all administrative tasks, at any and all levels. Stevens and Levi gloss this point somewhat, but they do mention in passing that metarubrics are most-often employed as a personal tool to objectively evaluate one's own rubrics and rubric use (also referenced in K. Schmidt, personal communication, Autumn 2012). Perhaps because they are pursuing a particular agenda in pitching rubrics as a program-level tool for individual students' learning assessment, Stevens and Levi may overlook the value of the metarubric in program (and maybe

even instructor) assessment. This is a significant omission, however, inasmuch as rubrics at all levels of application strike a useful balance between academic freedom and curricular/pedagogical standards (A. Myatt, personal communication, Summer 2012, S. Flaschka, personal communication, Autumn 2012), and this is arguably the most important takeaway that rubrics have to offer.

As to the particular tasks to which rubrics may be effectively applied, perhaps the most relevant to modern educational goals is the student *portfolio*, or its contemporary evolution, the student *electronic portfolio* (Banta, 2006). As a historically-collected set of performance task artifacts, the portfolio is emerging as the leading candidate to augment, or even replace, the series of conventional examinations that have traditionally been used to assess student learning. Interestingly, the value of a portfolio as an assessment technology lies – perhaps unsurprisingly – in the emergent phenomenological ontology that the artifacts of the portfolio shape. Banta also notes:

> No standardized exam is truly content free, and if it were, it would be a better test of general intelligence than of what is learned in college. The near-perfect correlation of CLA [the Collegiate Learning Assessment] scores with ACT/SAT scores suggests that the CLA may be a better measure of the abilities students bring with them to college than of the learning they take away. (p. 3)

Subsequently moving to a fundamentally qualitative perspective, the task for raters of portfolios then becomes one of evaluating the phenomenological ontology of the portfolio within the context of the desired leaning outcomes, rather than against some objectively-established content-based criteria. This perhaps daunting undertaking is therefore best attended to with a well-designed rubric that follows the guidelines I have set out above (following Terenzini, then

Stevens and Levi), approaching portfolio evaluation from the top down in terms of the need to assess student learning, but also from the bottom up in applied consideration of how those outcomes may best be captured in particular assessment criteria. I hasten to add however, that this shift in emphasis to looking at form-plus-content rather than content-by-itself for the purposes of qualitative assessment is not intended to suggest that quantitative, content-based, and even standardized, tests should be considered somehow obsolete, but rather that the recognized limitations of purely content-based assessments may be overcome using qualitative methods, and therefore a program may embrace and demonstrate a wider spectrum of learning outcomes – especially in the case of applied learning – than is reflected in the mere recall of discipline-relevant facts. If the portfolio, as we may have been promised, is truly a "golden door" that opens onto insights into the depth, breadth, and scope of a student's learning, then the rubric devised for assessing that portfolio may be thought of as the "golden key" that unlocks the door. This comprehensive functionality derives from the construct validity that forms an essential part of the assessment technology, as I will detail further below.

At this point, a concerned reader might raise the objection that the highly program-specific nature of a well-designed rubric effectively precludes such a rubric being of much use outside the particular institutional department where it has been devised and applied. This is indeed a meaningful concern, as the entire process of designing and refining a rubric to match the specific learning outcomes of a specific context would appear to make the instrument too individualized to have relevance elsewhere. However, this is precisely why I assert that rubric design should be approached in accordance with the systematic, theoretical frameworks built from the models of Terenzini and Stevens and Levi: by relying upon a common conceptual framework as a starting point, and presuming that there is a more-or-less objective nature to the

content of any given discipline, highly institution-specific rubrics may still share a common phenomenology (from the rubric-building process) and a common ontology (from the established conventions of the discipline); the differences between any two different rubrics used in different institutions (but similar or corresponding disciplines) will be, upon deep inspection, merely cosmetic, provided there is some inter-institutional correspondence between designated learning outcomes.

Thus, although the rubrics as instruments in and of themselves may not have transferability, the *process* whereby they are devised, applied, and revised will. In this *process-oriented* rather than *object-oriented* way, assessment that relies upon rubrics can serve the sometimes-incongruent goals of institutionally-individualized academic freedom and discipline-based-standards of academic accreditation *simultaneously*. This benefit derives from the methodology behind the process of consistent assessment practice facilitated by suitable rubric use, and not from any particular, individual details of how the process is manifested from institution to institution. In other words, how we assess is evaluated in terms of the inter-institutional, shared phenomenological ontology of the rubric-as-instrument, rather than a simple comparison of particular student performance tasks and their associated customized rubrics from one program to another. It should not be surprising that, given the emphasis on process (tasks that demonstrate applied learning, for example) that is typical of qualitative assessments in comparison to the purely objects-of-knowledge orientation (identification/recitation of facts, for example) that characterizes typical quantitative assessments, qualitative assessment can offer perspectives on student learning which are unique to qualitative technologies, but are universally-applicable to institutional, educational missions. Ultimately then, it is the

methodology of rubrics which is transferable, if the same cannot be said of any particular rubric in and of itself.

The Question of Validity

This perspective on transferability, underpinned as it is by a theoretical notion, rests heavily upon establishing the *validity* of rubrics and their associated phenomenological ontologies for its own legitimacy and virtue. In his landmark essay on the subject, Messick (1989) defines validity as:

> …[an] integrated, evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores. (p. 13)

I note that, especially for the present purposes, this definition should be construed as referring to an on-going process, rather than a static, one-off effort, and this makes it consistent with the continual re-evaluation process necessary for effective rubric use. Messick's definition is focused on quantitative measurements in the tradition Brown (1996) further defines as "the degree to which a test measures what it claims, or purports, to be measuring" (p. 231). Given that the history of assessment validity has developed a mature theoretical model that relies heavily upon *inferences* about evidence of learning, and even in light of the contemporary aphorism that "measurement is easy, meaning is hard," I contend that it is a straightforward (although non-trivial) task to insert considerations of *meaning* into validity theory along lines similar to those which have been well-established for *measurement*, and that the necessary groundwork for doing so has already been laid out by validity theorists.

In addition, as my analysis relies heavily upon the fact that the notion of *constructs* is central to modern validity theory, I propose some operational characteristics, extrapolated from

the literature, to conceptually delimit the concept of a construct here. Borrowing from behavioral psychology – and momentarily setting aside my focus on phenomenological ontologies – a construct may be simply defined as a *label* used to describe *behavior*. This plainly non-rationalist definition is purely operational, but given that as educators we must rely upon empirical, objective technologies when performing assessments (and can expect to continue to do so unless some form of telepathic technology is invented that allows us to directly access student thought processes to look for evidence of learning), the non-metaphysical, non-mentalist model from behavioral psychology is useful in practice. From this behaviorist perspective then, the label "construct" refers to a not-directly-observed (or in the psychological vernacular, "latent") characteristic of the subject (here, the learner). Typical examples of this are things like creativity, intelligence, reading comprehension, and beliefs.

The most important property of constructs that must be given due consideration is that, at least in the empirical sense, constructs do not have material-ontological status and thus do not technically exist in the objective, materialist sense of the word. This has given more empirically-inclined researchers a great deal of difficulty (Slaney & Racine, 2013), but it should be noted that empiricism has always had trouble handling *names* (since empiricists typically deny the pre-existence of the thing the name refers to, which can be highly problematical when dealing with real-but-non-material things such as *numbers*). This incidental point is not of much concern to my analysis, however; it is sufficient to take the label as behavioral psychologists prefer to use it at face value and, for the sake of analysis, allow it to have enough cognitive content to be taken as referring to something that emerges in the learner (either mentally or behaviorally, as one's philosophy may see fit) that is *constructed* as a result of the learning process. It is this useful and

informative thing, the "construct," that is then somehow to be measured by the assessment process.

In this way, it is not even strictly necessary to import the entire apparatus of a phenomenological ontology into consideration of a instrument's (here, a rubric's) validity; the process of evaluating validity may be performed from a purely-empirical (but perhaps not purely-materialist) perspective, and this point is crucial to establishing the objective nature of rubric-based assessments. Furthermore, the aforementioned reliance, detailed below, upon *inference* in establishing instrumental validity belies the inherent *mental* nature of the entire enterprise, empirical/materialist preferences, perspectives, and predispositions notwithstanding.

As originally considered by Messick (1980), test validity is theoretically conceived of as being a combination of three factors: 1) content validity, 2) criterion validity, and 3) construct validity. *Content validity* is concerned with the relevance and coverage of the domain under examination; *i.e.,* whether the instrument addresses the subject area specifically and representatively. This is the most straightforward of the three, but is included due to its foundational importance: measurements should not be so indirect as to lack meaningful relevance. *Criterion validity* has two dimensions: *predictive validity* and *concurrent validity*. Together, these two give the test its predictive and diagnostic utility (within the relevant theoretical framework under examination), and by extension, the test's *reliability*. However, by far the most complex and detailed of the three is *construct validity*.

Messick (1980) initially breaks this one down into a plethora of components: *convergent validity, discriminant validity, trait validity, nomological validity, factorial validity, substantive validity, structural validity, external validity, population validity, ecological validity, temporal validity,* and finally *task validity*. Ultimately however, drawing from his own research and that of

Cronbach (1969) earlier, Messick (1989) eventually moves to a unified conception of validity centered entirely around construct validity (Ruhe, 2002). This shift is made possible by further examination of what a construct is, and by subsequently extending and simplifying the definition of the term. Eventually, the unitary construe of construct validity emerges from the fact that there are many potential *sources of evidence* bearing on the appropriateness of inference(s) related to the construct of interest (Cizek, 1997).

In the modern vernacular of the American Psychological Association (APA) and the American Educational Research Association (AERA), "validity" refers to "the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores" (APA, AERA/NCME, 1985). Here, the traditionally-quantitative notion of "scores" is inclusively defined by Messick (1989) "broadly to mean any coding or summary of observed patterns on a test, questionnaire, work sample[,] or portfolio" (Ruhe, 2002, p. 149). This usage then explicitly extends the concept of validity to purely qualitative assessments, while simultaneously providing guidance into how to apply it.

Validation practice is the collection of evidence used to assess validity (Ruhe, 2002). Traditional sources of this evidence include "measures of content representativeness, internal consistency[,] and reliability and correlations with alternate measures" (Moss, 1992, p. 245), whereas in recent years, these evidentiary sources have been expanded to encompass both performance assessment and social consequences to further assess the functional worth of the assessment and by extension, the treatment/learning (Messick, 1998; Ruhe, 2002).

Messick (1998) eventually resolves that:

> All validity is of one kind, namely, construct validity. Other so-called separate types of validity – whether labeled content validity, criterion-related validity, consequential

> validity, or whatever – cannot stand alone in validity arguments. Rather, these so-called validity types refer to complementary forms of evidence to be integrated into an overall judgment of construct validity. What needs to be valid are the inferences made about score meaning, namely the score interpretation and its action implications for test use. (p. 37)

which, for my purposes, leads right back around to the importance of a good theoretical grounding for rubric design and use, rather than a mere mimicking of previously-effective rubric forms, to establish the parameters of *meaning* inside and outside the assessment context.

What eventually emerges from Messick's theorizing is popularly characterized as *Messick's Validity Framework*. This framework consists of a unified conceptualization of construct validity for test (I would prefer "technology" or "instrument," in the case of a task/rubric combination) interpretation and use, and has three component dimensions (Messick, 1995; Ruhe, 2002). The *evidential basis for test interpretation* is based upon the empirical analysis of all data used in construct validation and inter-construct relations. The *evidential basis for test use* also analyzes construct validation, but examines its relevance and utility to external, applied contexts such as social or vocational environments. Finally, the *consequential basis of validity* considers the value implications of score interpretation as a basis for action, and the actual and potential consequences of test use, especially as may regard sources of invalidity stemming from issues of bias, fairness, and distributive justice (Messick, 1980).

These last points require some further exegesis here. Modern educational theorists have extended this framework of Messick's to encompass at least two additional analytical perspectives (Moss, Girard, & Haniford, 2006). In addition to relying upon educational measurement in the traditional sense as set out by Terenzini, validity theory is nowadays also

informed by *hermeneutics* and *sociocultural studies*. On the one hand, considerations of validity as depending on an epistemological understanding of the relevant phenomenological ontology grounds that approach in a basic philosophy of (here, social) science that addresses the nature and justification of knowledge claims. On the other hand however, this basic assessment context may be extended by "interpretive social science" (p. 110) to provide further relevant insights into the nature and extent of learning.

Historically, validation practice has focused on the *intended* interpretations of test scores in lieu of deconstructing the test itself (Moss et al., 2006). A more flexible approach to validation practice should be one:

> …that can develop, analyze, and integrate multiple types of evidence at different levels of scale; that is dynamic so that questions, available evidence, and interpretations can evolve dialectically as inquirers learn from their inquiry; *and* that allows attention to the antecedents and anticipated and actual consequences of… [the learners] interpretations, decisions, and actions. (p. 111)

As validity theory evolved into a unitary notion of construct validity (Messick, 1989), it became increasingly important to understand the *value implications* of both score meanings and the use to which those scores were put (Moss et al., 2006). This is addressed in the first criterion of Messick's Validity Framework as he offers considerations of *sources of evidence* for basic forms of hypothesis testing in establishing the basis for test interpretation. These proffered sources are many, but Shepard (1993) notes that there are legitimate objections to the characterization of validation practice as pure scientific inquiry. For one, allowing considerations of score meaning to precede considerations of the use to which the test is put undermines the objectivity of the validation inasmuch at such a practice prejudices the contextualization process and biases the

application of any theoretical frameworks that might be appealed to in establishing validity. Furthermore, when this approach is combined with the ongoing process of validity re-assessment, such could incubate "the sense that the task is insurmountable" within the assessors which could in turn lead to shortcuts being taken in the rigor of the evidence-collecting practice (p. 429).

Although these concerns are addressed by the grounded-theory approach to rubric design and application that I am advocating, there is still room for extension and refinement. In light of the issues raised by Shepard in response to Messick's methodology, Kane (2006) calls for a less-abstract and more practical approach to validity theory built on two distinct, but complementary, kinds of arguments. First, Kane identifies the *interpretive argument* as one that straightforwardly sets out the various inferences and assumptions that lead from the observed task performance to the significance and meanings assigned to that task performance. From there, the *validity argument* uses logical reasoning and empirical evidence to evaluate the interpretive argument and each of its assumptions. Kane's idea is that by explicitly elucidating the interpretive argument in a separate process, it protects considerations of validity from making implicit – and therefore potentially-inappropriate – interpretive assumptions by making lacunae or omissions in the evidence harder to ignore.

Kane (1992) helpfully lists and describes the various categories of inferences used in building interpretive arguments:

- *Scoring (or observation) inferences* involve assigning some scaled, or at least relative, value to each performance or element of a performance
- *Generalization inferences* extend the interpretation from the observed performance(s) to similar tasks under similar circumstances

- *Extrapolation inferences* extend the interpretation even further from the generalized performance domain into a "trait" being assessed
- *Implication inferences* extend the interpretation to include verbal descriptions of scores, and claims or suggestions those descriptions might imply ("Average," for example)
- *Decision inferences* link scores to decisions or actions and their consequences (intentional or otherwise)
- *Theory-based inferences* extend to interpretation to address (presumed) underlying mechanisms, properties, and/or relationships that account for the particular of the observed performance
- *Technical inferences* address the appropriateness of assessor assumptions about technical issues such as the equivalence of performance forms, the extent of scaling, and the fitness of statistical models, techniques, and assumptions

Of these seven (or perhaps six, as Kane is occasionally inclined to fold implication inferences into decision inferences and/or technical inferences at various times in his discourse), the first three – scoring, generalization, and extrapolation inferences – are involved in nearly all interpretative arguments, following Kane's model (Moss et al., 2006).

Only with the interpretive argument firmly in place should we proceed to the validity argument, the purpose of which is to evaluate not the assessment *per se*, but the relevant interpretive argument *of* the assessment. Kane (2006) notes that some of the inferences of an interpretive argument may be taken for granted – such as "students can read the question" – but that others will need more careful evaluation – such as "the achievement test covers the content domain," which Kane duly notes is nearly always questionable (p. 23). As a hardline epistemologist and systems analyst however, I am inclined to question each and every

assumption equally. In my experience it is unwise to assume anything about student ability and preparedness; sometimes it is the case that students cannot in fact read (with the expected level of understanding) the question, and to that extent, precursory remedial pedagogy may be called for.

Nonetheless, as Moss et al. (2006) explain:

> Consistent with its heritage in a naturalist or unified approach to social science, validity theory in educational measurement supports the development and evaluation of interpretations based on standardized forms of assessment that are intended to be generalizable – meaningful and useful – across relevant individuals and contexts. Validity research is conducted, in part, to ascertain the extent to which such generalizations may be warranted. (p. 118)

Thus, if we understand the correspondence between the constructs grounding a particular rubric's discrete criteria and the interpretive argument(s) that are foundational to considerations of an assessment's validity, we have the basis for generalizing – and therefore implicitly also establishing the reliability of – an assessment technology. Reliability and generalizability may then be conceived of as being derivations of validity. I should note that this framework does not guarantee the converse dependency: reliability and generalizability do not in and of themselves yield validity. Reliability and generalizability may occur accidentally as a consequence of rigorous (re)design (as above in Harris, 1986); validity, however, essentially produces them, and therefore is more fundamental to the assessment process in general.

Returning to Messick's Validity Framework then, if the evidential basis for test interpretation is the starting point for the validation process, the validation process should

subsequently turn to address, in order, both the evidential basis for test use and the consequential basis of validity. Moss et al. (2006) note that:

> …validity inquiries are always situated within a particular social context and guided by the problem, issue, or question one is trying to address and the available resources (evidence, conceptual tools) for addressing it. (p. 129)

In addition to the aforementioned practices, this second-order process of *situated inquiry* examining uses and consequences may rely on *hermeneutics* and *sociocultural studies*, respectively.

Hermeneutics has its origin in textual analysis, but in recent years has been extended to cover pretty much any social phenomenon, including histories and cultures. "Hermeneutics is about the theory and practice of interpretation, about the bringing of understanding into language" (Moss et al., 2006, p. 130), and as such, it functions as a specific philosophical tool for analyzing phenomenological ontologies (as indeed might be expected, given that hermeneutics itself is an applied outgrowth of traditional phenomenology). Formally, hermeneutics is concerned with the "meaning" of a "text," but that analysis extends far beyond mere considerations of the denotative cognitive content in instances of subject-predicate grammar. As a second-order consideration, hermeneutics also examines the particular vocabulary, usage conventions, implicit assumptions, values, author voice, and community standards of word choices and uses, paragraphs construction, and rhetoric. As a third-order consideration, hermeneutics also considers the intended audience and the social context which a "text" is presented to and within. Without digressing into more-granular detail here, suffice it to say that the conventions of hermeneutics represent another vetting tool that educators can use in the validation process, if desired.

Likewise, sociocultural studies may be characterized as examining "relations among the person, activity, and situation, as they are given in social practice" (Lave, 1993, p.7). This broad definition may be concentrated in educational assessment to considerations of *learning environment*, *community of practice*, or *activity system* (Moss et al., 2006). Wertsch (1998) identifies the learning environment as encompassing "the relationships between human action, one the one hand, and the cultural, institution[al], and historical situations in which this action occurs, on the other" (p. 23). Lave and Wenger (1991) in turn describe (somewhat reflexively) a community of practice as "a set of relations among persons, activity, and world, over time and in relation with other tangential and overlapping communities of practice" (p. 98), noting that the analytical tools are "artifacts – physical, linguistic, and symbolic" (p. 57). Finally, Engeström (1999) identifies an activity system as using some form of mediated activity (such as an assessment scenario) to "explicate the societal and collaborative nature" (p. 30) of the actions under examination.

All of these theoretical perspectives represent different approaches to the same, central undertaking of ascertaining *meaning* in assessment scenarios. To simply the framework somewhat, Messick (1989) helpfully provides a general summary of the process and relations, crossing them in two dimensions he refers to as "facets" of validity inquiry (p. 20) and reproduced here (adapted from Brown, 2000):

Table 1. *Messick's facets of test validity*

	Test Score Interpretation	Test Score Use
Evidential Basis	Construct Validity	Construct Validity + Relevance and Utility
Consequential Basis	Value Implications	Social Consequences

The first facet (the columns) addresses the function/outcome of testing and therefore distinguishes between *interpretation* and *use* in the evidential basis for validity. The second facet (the rows) focuses on the justifications for testing, and therefore distinguishes between the appraisal of *evidence* and the appraisal of *consequence(s)* (Moss et al., 2006). In this matrix, *construct validity* is presented as defined by Messick (1989) to be the theoretical context of implied relationships to other constructs, whereas *relevance and utility* are defined as the theoretical contexts of (external) applicability and usefulness (*cf.* Brown, 2000). *Value implications* are defined as the contexts of implied relationships of performance scores to good/bad, desirable/undesirable, and similar judgments, whereas the *social consequences* are defined as the value contexts of implied consequences of test use and the tangible effects of applying the test. Messick notes that this matrix is progressive, in the sense that construct validity appears in each of the four cells, explicitly or otherwise. In this way, construct validity serves as the foundational, "integrating force" in validity inquiry (p. 20).

Given the highly context-dependent nature of even the most theoretically-grounded validation process, there is arguably no single best approach to evaluating construct validity. Brown (2000) gives a non-exhaustive list of suggestions that includes correlation coefficients, factor analysis, ANOVA studies, and mixed-mode studies, to name but a few. Since a program employing multiple analytical techniques to investigate construct validity can consume a significant amount of resources – especially time and money – is it fitting and perhaps ironic that assessment developers engaged in the validation process can benefit more from concentrating upon the quality of their inquiries instead of the raw quantity of such. The key takeaway here is that construct validation relies fundamentally upon assessors carefully drawing inferences –

making an "informed leap" in each case from an observed, measured value to an estimate of underlying standing on a construct (Cizek, 1997).

Addressing the application of all this apparatus of validity theory to assessment praxis, Moss (2007) then asks the crystallizing question: "…what is being validated: a test, or an interpretation and use of test scores?" (p. 473). Lissitz and Samuelsen (2007) object to Messick's argument that validity resides entirely outside of the test itself, concerned that Messick's approach essentially ignores the context within which the test/instrument is conceived and constructed – yet the externality of validity is a position that Cronbach (1971) also champions:

> One validates, not a test, but an interpretation of data arising from a specified procedure. A single instrument is used in many different ways… Since each application is based on a different interpretation, the evidence that justifies one application may have little relevance to the next. Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test "is valid." (p. 447)

In this view, *the test itself is not centrally important*, as it is merely a tool to be used. *How* and *why* the test is used, and *to what use it is put* are the far more salient questions for validity theory. Therefore, appeals to the circumstances of an instrument's source are considered to be either a form of *red herring fallacy* that diverts attention from consequences to origins, or a form of *false cause fallacy* that asserts an incorrect origin for the phenomenon (*i.e.*, the validity of an instrument is in truth not specifically contingent upon its origin). This is a subtle point, and it has generated significant debate in the literature. As I see it, the confusion arises from the fact that on the one hand, it is essential to systematically approach the design of an assessment instrument in order to instill a predilection for yielding validity (and consequently reliability) in it, but on the other hand, validity can only be ascertained by *a posteriori* analysis and examination of the

instrument's application. This, in my view, is precisely what drives the need to continually re-evaluate qualitative instruments; no instrument (qualitative or otherwise, but especially qualitative) ever springs forth *a priori* fully-formed and perfect from its creator(s), and validity is, as a matter of practice, only ascertainable empirically and externally to the instrument, as part of the overall technology of an assessment program. The methodology I have presented above is intended to facilitate the development of valid qualitative instruments *by design*, but it can never effectively guarantee the validity of an instrument (here, a rubric) as such; the validation process will always require appropriate reflection and analysis after the fact. I am merely looking to streamline the process and not only reduce the labor input required in the assessment process, but also avoid what software engineers used to refer to as "Garbage In, Garbage Out" results in assessment programs. In effect, instead of offering simple checklist, by digging into the underlying theory and analyzing it with an eye toward application I am asserting that a systematic approach – as complex or simple as may be deemed appropriate by the stakeholders – is necessary for both success and efficiency in the assessment process, especially when working in the realm of the qualitative.

Other Practical Considerations for Reliability

Beyond a firm and fundamental theoretical grounding in validity theory, reliability requires – in addition to an empirical, field-test-based program of revision – a well-designed and well-maintained process of *calibration* (Chun, 2010). In the context of assessment, calibration is an arguably unique criterion in that it is focused on those who are doing the assessing, rather than those who are being assessed or even the instrument used to assess them. Calibration provides the other essential component of reliability, as it provides a check, external to the assessment task the subject is performing, against the normative values the instrument scores against.

For example, although a consistent phenomenological ontology is a presumed feature of a properly-devised and properly-revised rubric, calibration is necessary to ensure that different, individual raters apply that rubric consistently between different test subjects and over different groups and sessions (A. Myatt, personal communication, Spring 2012). Even experienced raters can become "rusty" within a short time frame (a week is often sufficient to degrade rater performance), and a few simple calibration exercises prior to a session of rating can quickly normalize the practice of scoring, whether raters are using a scoring guide rubric or a more-detailed multi-level rubric (S. Flaschka, personal communication, Spring 2012).

Normalization from calibration is a necessary part of qualitative assessment applications to ensure scoring consistency absent the presence of a standardized quantitative structure to a rubric, yet the literature to address this issue is surprisingly sparse. On the one hand, this may be due to the implicit assumption (perhaps a typical one) that providing a scoring scale within the rubric itself is sufficiently-quantitative to guarantee a normalized structure to the assessment, yet on the other hand, as construct validity is not an explicitly-objective entity and requires empirical examination to confirm its presence, we have no unequivocal reason to expect such a guarantee to arise necessarily. Furthermore, given that it is human nature for an individual's cognitive performance to vary over time as a result of both internal and external factors, prudence dictates that we take such variability into account when we undertake a rating session.

It is this variability (or less charitably, "fallibility") among human raters that has driven recent research into using automated computer systems, employing various schemes of artificial intelligence and/or pattern recognition, to attempt to supplant humans in the role of raters/scorers for assessment programs. Unfortunately, these attempts to mechanize the process have yet to really demonstrate reliability (Stross, 2012; S. Flaschka, personal communication, Spring 2012).

Without delving into all the technical details, it is sufficient to note that the current approaches to automating qualitative assessment focus primarily upon textual essays and tend to rely upon *statistical modeling* of the text to evaluate the likelihood of the content being cognitively meaningful. Unfortunately, despite the enthusiastic optimism of its proponents, this approach is still easily "gamed" due to the inherent cognitive limitations of the model (Winerip, 2012). The problem arises not from the fact that the statistical modeling itself is somehow unsound, but from the fact that language, as per the theoretical perspectives mentioned above, is often (and easily) highly-charged with what is colloquially known as "meta-content," in reference to all the contextual elements that give language (and language use) its full *meaning*, in the cognitive, phenomenological, psychological, sociological, circumstantial, political, and construct validity sense(s) of the word. For illustration, I have included a landmark example of this problem, recently written by Les Perelman of the Massachusetts Institute of Technology, as Appendix C. Statistical modeling by itself alone is purely top-down, and cannot address the fundamental bottom-up requirements of construct validity as I have laid them out. As mentioned earlier in this essay, statistics can yield only measurement, not meaning, and some additional apparatus of a phenomenological nature is required for proper (valid, reliable, useful) qualitative assessment: also as above, the analysis of content is a separate and more-complex task than the analysis of form.[5] Statistics are not, by themselves, inferences, and unless/until technologists can produce a

[5] This is a case of a "P versus NP problem" for computer science. Since construct validity can *verified* in polynomial time (by humans), the unanswered-at-present research question for software engineers and computer scientists is whether or not construct validity can be *solved* in polynomial time (by a Turing Machine algorithm, for example). For further background on this issue, see the historical survey in Auerbach (2012).

cybernetic, mechanized system to undertake this *additional* dimension of assessment, it will still fall to humans to supply the contextual analysis that gives insight into meaning, truth, and learning.

There are more practical alternatives to automated rating, however, to ensure calibration's *inter-rater reliability*. Whereas specific measures of IRR may be quantitatively analyzed using statistics such as *Cohen's kappa* (in the case of two raters) and *Fliess' kappa* (in the case of two or more raters), or the broader *Krippendorf's alpha* (comparing coded data to rated scores), the important point here is that such measures should be made possible by any given assessment program in the first place. In addition to having a basic, grounded validity, a qualitative assessment technology will provide the highest reliability when multiple raters assess the same performance task of the same subject(s) (Krippendorf, 1970; Krippendorf, 2004). Experience – and the existence of Cohen's kappa as a distinct calculation separate from Fliess' kappa – indicate that pairs of raters can provide an optimum trade-off in labor costs (typically, time rather than money) and scoring consistency (*cf.* López, 2002). Thus, raters performing qualitative assessments should ideally be grouped into randomly-assigned, then randomly-reassigned-often scoring pairs, as this will tend to preserve the initial calibration of a rating session for the duration of that session (S. Flaschka, personal communication, Spring 2012).

This integrated approach to design, usage, and evaluation ultimately works because the reliability of a rubric lies not in the instrument itself, but in the raters and the inferences involved in its usage.

Concerns are sometimes raised and debates are sometimes sparked among assessors over the question of whether or not to provide students with access to a rubric before the students attempt the performance task. At issue is the potential problem of students specifically tailoring

their performances to meet the explicit expectations delineated in the rubric's scoring levels. As I see it, the existence of such a perceived problem may be taken as being more indicative of some shortfalls in the alignment between the instrument and the intended learning outcomes than of a difference of perspective on pedagogy. If proper content validity has been pursued in the design and evolution of a rubric (as above), then what is really at issue here are the purposes to which the assessment/instrument is being put.

Although there are well-established pedagogical schools of thought that emphasize discovery processes and educational constructivism within the learner (which I endorse, as above), telling students what we want them to learn should not invalidate the assessment process; if it does, then the construct validity of the assessment instrument may be suspect, and revision of the rubric in use – or even the learning outcomes – may be called for. Only in cases of rigorously-applied Socratic Method do we find pedagogical scenarios and circumstances that leverage keeping the students in the dark at the outset into advantages for the intended outcome. Indeed, when students have the opportunity to provide feedback to those who have taught and assessed them, one especially salient, emergent theme is that of seeking to understand the expectations teachers have of them as students (Stevens & Levi, 2005; OECD, 2010).

Given the primacy of student-teacher relationships to the efficacy of the educational enterprise, we should be careful of undermining the essential trust in one's mentor(s) that such relationships rely upon; as educators, it is generally not our purpose to conceal things from our students, except in those cases where self-driven-discovery is part of a pre-identified learning outcome. As assessment *per se* is not in and of itself a learning outcome, the only purpose concealing assessment criteria beforehand can possibly serve *is another, perhaps tacit learning outcome*. This second-order outcome is then the proper subject of a second-order assessment

once the second-order outcome is revealed. Theoretically, this compounding process could be repeated *ad infinitum*, but we would eventually expect the students to learn anticipate it (*cf.* "enlightenment") and thus rob it of its effectiveness (which, of course, could serve as yet a third-order learning outcome). In this way, when done correctly, initially obscuring the assessment criteria as part of a pedagogy derived from the Socratic Method can yield paradigm-shifting insights in the learner; overusing assessment-as-content-delivery, however, risks such a technology becoming detrimental to student engagement.

Leadership and Management Considerations

Although I have not made the role of *leadership* explicit in the previous sections of this chapter, as I now move from an examination of the conceptual and measurement issues to turn my attention toward the organizational and political ones, as a practical matter, its role must now become central. From the organizational (and by implication, the institutional) perspective, the major issues for programs of qualitative assessment come down to ones of *cost* (Ewell & Jones, 1985) and *culture* (Shipman, Aloi, & Jones, 2003), and addressing these issues necessarily expands the considerations more broadly outward from applied instructional practice, and thus embraces the more-immediate involvement of institutional administrators at various levels with the assessment process.

Barely one year after Ewell (1984) offers assurances that assessment does not require a "massive" increase in costs (as above), he acknowledges (Ewell & Jones, 1985) that there will be *incremental costs* involved in implementing updated, modernized assessment programs. Specifically, amid calls to "measure your mission" (p. 6) administrators of assessment programs must successfully address different cost considerations, some of them common to all

administrative programs (especially educational ones) and some of them unique to assessment processes (especially student-centric ones).

The most obvious of these varied concerns are the *direct costs* such as those associated with deploying a new assessment instrument and/or technology for any given assessment project, as well as those associated with the analysis of the results of that project. As with most itemized costs, these are largely inescapable, yet administrators need to properly anticipate and control them. It is this aspect of assessment projects that my analysis is most closely aimed at in the end; my theoretical model of what makes for "better" qualitative assessment is intended to facilitate a framework of methodological understanding that in turn allows assessors to realistically estimate the time, labor, and material that will be required, especially in contrast to legacy and/or traditionally-qualitative assessment programs. One of the key points that Ewell and Jones stress is that this cost represents a *regular, ongoing investment*, and not a *one-time expenditure*, and therefore assessment must be conceived of and understood as a continual process that adds value to the institutional mission, and not as an occasional, only-as-externally-required activity (p. 7).

Hand-in-hand with the direct costs are the *indirect costs* of implementation, and these are largely dominated by matters of overhead. Although Ewell and Jones (1985) do not address it in much detail, implicit in the framework of "overhead" is the notion of *opportunity cost* that may be measured in terms of faculty and administrative professional time spent on the tasks of assessment instead of other service/teaching/research activities. This investment choice extends to both the time spent on the analysis of specific assessments themselves as well as time subsequently spent on program reviews and revisions as a consequence of the results of assessment analysis. There is, not surprisingly, no straightforward, tested, theoretically-grounded formula appearing anywhere in the literature that gives either a simple rule of thumb or a precise

calculus of how to establish the parameters of either of these cost issues in an assessment-specific context, and it is unclear if such a thing would even be useful; in practice, budgeting is more of an art than a science, and although each institution must weigh the costs and benefits of assessment for itself and in light of its own mission and intended outcomes, if the factors are well understood, the resource needs may be reasonably-well anticipated (Ewell & Jones, 1985).

For planning purposes, costs may be considered from three perspectives, each with a different analytical purpose. *Full* cost represents the total project requirements for all stages and elements, and is the last perspective to consider, as it directly affects budgetary concerns. *Average* costs represent per unit costs (*e.g.*, per student, per faculty member) within the context of higher-order units, and typically inform variables used to calculate full costs. The most important perspective, however, are *marginal* or *incremental* costs (Ewell & Jones, 1985). Marginal costs depend upon the unit of analysis (per student, per faculty member, and here, even per program) much in the same way that average costs do, but are far more important to forecasting, as they may vary non-linearly. For example, economies of scale may provide very small marginal increases in cost when a technology is more-widely employed – once the host platform is installed and being maintained in operational condition, storing 100,000 student electronic portfolios will cost only slightly more than storing 100 of them – whereas the amount of staff time required to assess a large body of task outputs may incur increasingly-significant opportunity costs as other activates/responsibilities are progressively neglected – a faculty member can effectively and reliably grade a dozen or two term papers in a weekend without falling significantly behind in other duties, but a department will be unable to retroactively assess all writing samples collected from all sections of a core-curriculum course in the last five years and still balance a full semester's academic workload without some outside assistance. A clear

understanding of the complexities of these matters and their consequences for resource allocation is essential to exercise effective administration over them.

To assist in that understanding, for accounting/budgeting purposes cost elements may be grouped into four distinct categories for consideration (Ewell & Jones, 1985). First among these are the costs of the *instrument* itself. In qualitative assessments, instruments are usually developed from within the institution (as per the guidelines I have set out above), rather than being externally licensed. As such, and as with all of these elements as applied to qualitative assessment programs, the costs tend to be dominated by considerations of the measure of faculty time involved. Unique to this particular element however is the fact that qualitative instruments require an ongoing investment of (human) resources to operate and maintain, given the necessary process of continual evaluation and revision that they require in comparison to the convenience of prepackaged, conventional quantitative, externally-acquired assessment instruments that typically incur either one-time or (occasional, depending upon licensing details) fixed-recurring costs.

Second in this consideration are the costs of *administering* the assessment. In the case of internally-developed instruments such as rubrics employed in a performance task assessment, this cost is fairly linear when scaled, and is usually allocated on a per student basis. For present purposes, it is most telling and germane that Ewell (1984) recognizes that all instruments intended to assess critical thinking are *local* in origin, as they are highly dependent upon context and curricular delivery.

Third are the costs of the *analysis* of the assessment, and this is the point where costs require the most management (Ewell & Jones, 1985). Analysis encompasses both the study of student performance with respect to the intended outcomes as well as examining the reliability of

the technology and considerations of potential revisions to the instrument and/or both the assessment and the instructional programs. Without effective management toward a clear set of goals, these activities can consume an ever-increasing, uncontrolled amount of resources, as well as produce diminishing returns on resource investment, and combined, this can constitute a highly-undesirable program outcome.

Fourth are the costs of *coordination* among all the program and institutional elements involved, and this cost is often the most underestimated. Again, insightful management is called for, as this element perhaps most of all is rife with potential for wasteful expenditures. To this point, Ewell and Jones recommend several administrative tactics to control the various costs associated with any assessment program that encompass all four of these budgetary elements.

At the outset of any assessment program, it is useful to begin by taking an *inventory* of what assessment data may already be already available on students, programs, and the institution, in order to avoid any needless duplication of previous efforts (and, I would add sanguinely, needless repetition of previous mistakes). This pre-existing assessment data may have been taken for a variety of (other) purposes, but is often not centralized nor indexed for easy availability. Studying this data can inform not only the administrative planning of the new assessment program, but can also shape its ontological structure, its sought outcomes, and its design and implementation details.

In addition to being a goal of the inventory process, good administrative oversight of all parts of an assessment program avoids *duplication* of effort, particularly between different institutional departments and divisions (or, in the vernacular, "units"). Although a bottom-up approach to qualitative assessment requires a decentralized structure to function correctly (as explained above), it is the proper role of centralized, top-down administration to maximize the

efficiency of concurrent programs across the institution. Modern leadership theory (discussed in more detail below) suggests that *facilitating communication* is a more-effective strategy in pursuit of this goal than taking a directly-controlling management role can be. The previously-mentioned communicative value of rubrics is highly germane to facilitating this course, and can therefore provide an essential tool of intra-institutional (and inter-institutional, in accreditation contexts) consolidation of resources and data.

Complimentary to these two preliminary undertakings, looking for *mutually reinforcing information* is an essential task of program oversight, and one that may need to be driven by institutional administrators absent a vested interest by departmental-level assessors. This function is essential to the higher-level assessments of programs and institutions, but may also inform the revision of assessment technologies at all levels.

Finally, Ewell and Jones (1985) note that "Careful *tailoring* of data collection to instructional mission can limit costs" (p. 31, emphasis mine). Limiting the scope of assessment programs largely to specifically-identified needs prevents overtaxing the finite (human) resources of an institution, especially at the expense of scholarship and educational delivery.

Parallel to this budget-centric model, from the leadership perspective (focusing on principally personnel considerations rather than principally financial ones) Banta and Blaich (2011) characterize the process of effective outcome assessment programs in terms of three successive stages: *planning*, *implementation*, and *improving and sustaining*. In the initial planning stage, good leadership involves all stakeholders from the outset of the program in order to address their respective needs/interests in exchange for their later support of the program.

It is important for the planning stage, which should begin once the need for assessment is formally recognized, to allow sufficient time for a plan to be developed that has clear purposes

articulated and directly relates to the goals of the respective stakeholders. Establishing well-defined, unambiguous, and explicitly-stated program goals is a necessary prerequisite to program success when large and diverse interests are involved. This is not to say that assessment programs cannot necessarily have open-ended goals, but the expectations for and values of such goals will need to be well understood by all concerned if they are to be realistically pursued. As a noteworthy aside regarding the specific issue of faculty involvement in assessment programs, Ewell mentions elsewhere that leaders should "remember that you don't need everybody on board to move forward" (Hutchings, 2010, p. 3), and although this may be true in so far as it goes, prudence dictates that consensus is still highly desirable even when it is elusive or impractical (due to issues of resistance based upon individualized notions of academic freedom, for example; more on this below), and in practical matters of non-faculty externalities such as appropriation/grant/funding requirements, legislative mandates, governance directives, and the like, consensus with, and even among, external stakeholders may be an imposed prerequisite.

Banta and Blaich (2011) most extensively emphasize the role of leadership in the implementation stage of an assessment project. It is in the implementation that the details and complexities of an assessment project pose the most challenges to its success, and where maintaining the engagement of the various stakeholders – particularly the faculty – can be most difficult. To this end, helping everyone involved remain mindful that assessment is central to learning and is therefore in everyone's interest, providing professional development opportunities for involved faculty and staff, and providing good communication among the various stakeholder constituencies can all become crucial to the efficiency and efficacy of the program. Likewise, good leadership also attends to the practical, technological needs of the assessment, placing responsibility at the unit (typically, departmental) level (and consequently

allowing units to exercise that responsibility) while also seeking to facilitate sufficient validity and reliability of the assessment (thereby ensuring its utility). Throughout this stage, leadership should likewise facilitate the assessment of processes along with that of outcomes, so that outcome assessment can also serve as program assessment.

Finally, leadership is essential in the ongoing process of improving and sustaining an assessment program, in terms of both the assessment technology itself, as well as that of the overall program. Even in cases where a program generates a large amount of enduring momentum internally without continual external motivation, leaders need to nurture even that process.

The single most important factor for leaders to address in assessment programs is *engaging the faculty* (Banta & Blaich, 2011). Governance issues notwithstanding, faculty engagement is the cornerstone of the entire assessment process, and without nurturing and securing it, assessment programs simply cannot be effective (Terenzini, 1989). Specifically, "If faculty do not participate in making sense of and interpreting assessment evidence, they are much more likely to focus solely on finding fault with the conclusions than on considering ways that the evidence might be related to their teaching" (Banta & Blaich, 2011, p. 24). *Engaging the students* follows as a close second (Banta & Blaich, 2011). Many, if not most, assessment programs are driven by accreditation concerns, and students, who are often already challenged by the need to engage with their institutions, may feel even less inclined to engage with external mandates presented to their institutions.

Less directly-relevant to issues of leadership, but even more important to my analysis, is the historically *high turnover rate in faculty and administrative leadership* that assessment programs are prone to (Banta & Blaich, 2011). Banta and Blaich report (or perhaps "lament")

that "using evidence to promote improvements is not yet a core institutional function" (p. 26). As a consequence of this "churn" (again borrowing from the common vernacular) in leadership, assessment programs that are both long-running and successful are extraordinarily difficult to find examples of. This, finally, strikes at the root of the problem I pointed to in earlier chapters: assessment – particularly formative and summative qualitative assessment – *must* be improved upon somehow, even in light of internal and external factors that cannot be directly countered.

Furthermore, although Banta and Blaich (2011) are somewhat coy in addressing it, experience confirms that high turnover rates in leadership within the institution combine with short election cycles outside the institution to create immense political pressure on institutions to demonstrate short-term gains, even at the expense of long-term ones. The resulting *unrealistic timelines for change* driven by a persistent external insistence on "accountability" within the institution can lead to faculty frustration, resistance, and even outright rebellion (as above). This response in turn then fosters stronger calls for accountability as the next election cycle looms, and thus sows strife within the broader community of stakeholders, resulting in an increase in conflict rather than in communication.

It should be noted that it is also possible for assessment program administrators to lack the necessary professional confidence or decisiveness to implement assessment-indicated changes in curricula, pedagogy, and/or departments based on a (mis-)perceived lack of sufficiently ample and reliable data upon which to base expensive and perhaps unpopular decisions (Banta & Blaich, 2011). Sometimes, as a consequence of intra-institutional 9and even intra-departmental) politics, administrators may find it easier to suggest "further study" of a problem than to actually solve it – especially when one will be moving on in a couple of years anyway, and would therefore wish to avoid needlessly antagonizing one's colleagues by

implementing unwelcome, assessment-driven changes to the *status quo*. This is not a prescription for a successful assessment program outcome, as it fails to address and resolve what is perhaps the single most formidable obstacle an assessment program may face.

Ewell (1984) anticipates this, however, and identifies two distinct motivations for faculty resistance:

> A first reason for resistance is a fear on the part of the faculty that they will be negatively evaluated. A second basis is more philosophical: a conviction that the outcomes of what they do in the classroom are inherently unmeasurable by anyone but the faculty. (p. 78)

Ewell goes on to explain that part of the first problem stems from the tendency of the faculty to confuse/conflate assessments of program effectiveness with course or teaching evaluations – instruments that are widely considered to be highly prone to biases that make them unreliable. Therefore, for leaders in either the faculty or the administration, it is essential to place the focus of an assessment program clearly and unambiguously on the *curriculum* itself and nowhere else, involving the faculty intimately and from the outset by communicating and sustaining this goal to them and with them.

Ewell (1984) also notes that

> Faculty resistance based upon fear of negative evaluation is often heavily bound up with more basic objections to measurement of any kind. Many faculty are simply philosophically opposed to explicit outcomes measurement. They feel that it is inherently misleading, oversimplifying, or inaccurate. Moreover, many faculty believe that assessments designed to tap general attributes do not adequately reflect the specific emphases that the feel are present in their classrooms. (p. 79)

Here, then, the role of qualitative assessment through the mechanism of construct-valid rubrics may find its fundamental utility: providing *meaning* to learning and learning outcomes instead of mere *measurement* of them.

Ewell (1984) observes that teaching and scholarship are "fragile" practices that operate at their best within a decentralized and values-based context, insulated from an overarchingly instrumental environment (p. 13). In addition, he specifically asserts

> …that to achieve excellence in the diverse activities currently comprising postsecondary education, we must create explicit, institution-specific mechanisms for regularly assessing the degree to which we are in fact attaining our collective goals. (p. 13)

Further on, Ewell consequently notes

> The challenge to [the] administration… is to create explicit, information-based structures of incentives and accountability to replace our more traditional implicit methods of self-assessment and self-improvement. (p. 15)

Ewell's subsequent claim that the effectiveness of assessment programs "is highly dependent upon their being institution-specific and participatory in character" (p. 17) is much more in line with recent experiences reported in the literature. Ultimately, I think this last quote reveals that Ewell implicitly recognizes the necessarily "bottom-up" quality the assessment must have, even if he is couching his presentation from the administrative-managerial perspective.

Faculty Matters

However, as alluded to previously, a major critique of higher education faculty, dating back to at least the 1970s, is their widespread, determined – even obstinate – resistance to change (Tagg, 2012). Tagg identifies a double standard, "in which quality only becomes a question when contemplating change and no comparative evidence ever emerges" (para. 6) that is often

used by faculty committees to terminate proposed revisions by citing undocumented concerns about maintaining a nebulous notion of "educational quality." Thus, in perfect irony, preserving the assumed quality of a curriculum or pedagogy is often cited as a reason to oppose a new-and-therefore-somehow-disruptive formal assessment of the quality of that particular curriculum or pedagogy. Naturally then, aspiring reformers should always couch their issues in terms of assessing *how teachers can help students learn better*, for this *framing* tactic is difficult to oppose in purely abstract or non-evidential ways.

Tagg (2012) advises would-be reformers that, because of the natural human proclivity toward biological and psychological homeostasis, change most often comes in response to external stimuli. Therefore, the implementation of a process of *designed change* – a change that deliberately alters the rules of an activity for some specific purpose – becomes necessary for facilitating reform (para. 12).

Tagg (2012) also observes that developments in psychology over last the one hundred years or so have led to the modern understanding that people do not make choices from a purely-rational analysis that seeks the optimum outcome (*i.e.*, John von Neumann and Oskar Morgenstern's *expected utility theory*), but rather that people make choices based upon highly-subjective perceptions of *value* (*i.e.,* Daniel Kahneman and Amos Tversky's *prospect theory*) that are conceived of as relative *gains* or *losses* as measured from some chosen or provided *reference point*. Though I note that this process is especially prone to logical fallacies and therefore should not be considered reliable, it is unfortunately also a *descriptive* phenomenon in that is an observed default behavior in humans. As a direct consequence of this behavioral phenomenon, how choices are framed in terms of potential gains and losses can introduce a significant bias into how the decision of which choice would be preferable is arrived at. Whereas

semanticists might quibble over the specific phraseology used in some of the textbook examples provided by Kahneman and Tversky (2000), the general finding that people are risk-averse when choices and outcomes are presented in terms of gains, but risk-seeking when the exact same choices and outcomes are presented in terms of losses has been well-documented in subsequent research (see Tagg, 2012, for further background).

The leadership implications for educational reformers are then clear: faculty are only human, and presenting assessment programs to faculty as an opportunity for institutional gain is a recipe for failure, as naturally risk-averse mindset is likely to take hold within the faculty, and thereby give rise to a defense of the *status quo* no matter how unfounded such a defense might actually be. Once this mindset is in place, at least for American subjects, the *Dunning-Kruger Effect* can also come into play (Kruger & Dunning, 1999), causing the subjects to cling to their mindset even more stubbornly when presented with contrary evidence, out of fear of admitting (to themselves and others) that their initial reasoning was erroneous and of consequentially losing social and political capital within their peer group from this show of weakness.[6]

[6] There is some evidence reported in the literature of the Dunning-Kruger Effect being *reversed* in Asian cultures. In such cultures, social and political capital is often gained by promoting group harmony rather than by asserting dominance over the group, and therefore the sometimes highly-ritualized public admission of mistakes and errors – often requiring prescribed acts of contrition – is considered virtuous and valuable. See DeAngelis (2003) for more details, and note that the cultural implications of this difference extend to many leadership contexts, but that the main point regarding biases addressed from prospect theory and framing still holds across all studied cultures.

In addition to this cognitive and social obstacle from prospect theory, Tagg (2012) identifies another psychological effect that operates as a corollary of loss aversion, pointing to what Thaler (1980) describes as the *endowment effect*. In essence, the endowment effect operates to tacitly and subjectively assign a higher value to things which the subject already possesses, further driving risk aversion even when risk may not be present, and in the extreme causing some subjects to embrace a net loss by choosing to keep what they have possession of even when they can be guaranteed to obtain something of greater objective value in exchange for it (Ariely, 2008). Together, loss aversion and the endowment effect add up to what Tagg (2012) labels *the status quo bias*, borrowing from Kahneman and Tversky (2000) to characterize it as a perceptual illusion rather than a computational error. Samuelson and Zeckhauser (1988) suggest the status quo bias results from an attempt to resolve the *cognitive dissonance* that may surround estimations of one's own worth as a decision-maker, especially in light of the psychological need to justify past decision-making and resolve indications of past errors with the present self-estimation of oneself as a good decision-maker. Thus, Tagg (2012) identifies the status quo bias, which can occur at any level of an administrative hierarchy, as the principle obstacle for designed change a the unit (here, the departmental faculty) level.

Summary

All the component pieces for a grounded theory to improve qualitative assessment programs are now in place. From Terenzini, I have defined and delimited the fundamental parameters of assessment that form both its conceptual and pragmatic foundations. From Ewell, I have focused this framework upon modern educational contexts. From Stevens and Levi, I have qualified and examined the instrumental issues qualitative assessments face, as well as how best to address them in the integrated technology of rubrics, and from Messick I have delineated the

necessary basis for establishing the construct validity of those rubrics. I have subsequently looked to Kane and Moss *et alia* to further characterize the relevant inferences that may be drawn to give insight into construct validity within the context of institutional mission. Then I have looked to Ewell and Jones to examine and understand the nature of the costs associated with assessment programs. Finally, I have turned to Banta and Blaich to describe and explain the significant leadership issues that assessment programs face.

In the next and final chapter, I will bring all this together, with some further analysis and some additional considerations, into my conclusions and recommendations.

CHAPTER 5: ADDITIONAL CONSIDERATIONS, GENERAL CONCLUSIONS, AND SPECIFIC RECOMMENDATIONS

Overview

We should never lose sight of the fact that institutional change is the implicit goal of assessment (Terenzini, 1989). To guide the institution of such changes, a general series of recommendations for leadership in assessment programs has emerged in the literature, both in the broader context of institutional management in any commercial or public scenario, and within the specific context of higher education. The checklist provided to postsecondary educators by Jones, Voorhees, and Paulson (2002) is a typical guide of this type, derived from a mixed-methods study relying upon interviews and statistical analysis and sharing a large commonality with the general professed wisdom on these matters, and so I shall use it as an exemplar.

Jones *et alia* (2002) describe successful assessment leaders as

- being directly involved in the assessment process
- meeting regularly with assessment personnel
- facilitating communication
- establishing mutual trust
- promoting collegial collaboration
- providing real incentives for program participation and support
- fostering a deliberate planning process

- pursuing slow, incremental changes to maximize the potential for success
- integrate assessment and budgetary concerns

I note that none of these are particularly new to my analysis, but the crucial factor is that, as is usual for this type of advice, Jones *et alia* support these conclusions from the perspective of trial-and-error "best practices" rather than a grounded theory.

By examining the details of how and why qualitative assessment works the way it does, considering the established challenges assessment programs pose for their leadership, and applying some systems theory I can now provide a more-specific phenomenology to facilitate qualitative assessment program success.

Motivations

As leaders, when considering how best to motivate others, we should always take into account the others' *enlightened self-interest*. Even the most altruistic of faculty members, the ones most inclined to serve the public good of education even at significant personal cost, cannot be expected to abandon the well-being of their careers completely. Central to this is the feedback loop that has emerged in higher education wherein faculty who devote more time to research and publishing are paid more than their teaching-oriented colleagues, and this brute fact holds across the entire spectrum of postsecondary education (Fairweather, 1996). The feedback loop arises because this discrepancy in emphasis inevitably has significant consequences for tenure and promotion decisions (Schuster & Finkelstein, 2006), and thus a classic *vicious circle* forms wherein teaching is systematically and continually devalued as a professional activity among faculty in higher education. Tagg (2012) notes that "teaching load" has even begun to take on a pejorative meaning, in that faculty increasingly view time spent on instruction as a net loss in time that could otherwise be spent on research.

The only way to offset this shift is for administrators to pursue initiatives that emphasize excellence in teaching as being highly-relevant (or for the truly revolutionary, central) to tenure considerations. There are a few scattered examples of such initiatives that can be found referenced occasionally in the literature, but there is no evidence of widespread emphasis on such reforms trending across the overall spectrum of higher education institutions; to date, such initiatives appear to be the exception, rather than the rule (*cf.* Huber & Hutchings, 2005). This is the issue that administrators must address first and foremost if any assessment program – but most particularly a labor-intensive qualitative assessment program – is to be successful, or even possible as a practical consideration. All other issues regarding the improvement of qualitative assessment practices pale in comparison to the significance of this one.

In order to address this problem directly, it is necessary for administrative leaders in higher education to confront the settings wherein the mechanisms of tenure and promotion decisions principally reside in the loosely-coupled hierarchy that characterizes modern institutions of higher education: the discipline-oriented departments. Discipline-centric departments represent a double-edged sword academically; it may be viewed as scholastically advantageous that a community of researchers sharing a common professional specialization can work together under the appointed leadership of one of their peers toward the preservation and expansion of their particular, chosen body of knowledge, but emphasizing the necessary maintenance and advancement of their discipline and thereby themselves typically works against the faculty fostering a general public competency in (or even a serviceable familiarity with) that discipline through the mechanisms of education. Given the typical resource constraints under which faculty must function, and the self-reinforcing emphasis on publishing under which

faculty must survive, an "either/or" dilemma is created for the faculty in terms of how best to apply their energies, and the resulting choices are predictably discipline- and career-centered.

A further complication to this situation is that higher education faculty may often lack what I would describe as "sufficient" teaching experience, and may have only rudimentary pedagogical skills. This can often be the result of graduate education programs that emphasize the core competencies of research and publishing while leaving candidates to their own devices in terms of developing instructional proficiency. This emphasis then becomes self-reinforcing as successful academics subsequently carry this value system with them from graduate school to new jobs at new institutions, where they pursue and are eventually granted tenure, preserving not only the core knowledge of their disciplines, but also the traditional, scholarship-prioritizing value system.

Ultimately, Tagg (2012) describes the prevailing wisdom in departments regarding pedagogical improvement as "a morass in which gains would be invisible if achieved and in which [faculty] can only lose time, money, and energy through vaguely configured efforts that create no enduring value" (*circa* para. 47). I assert that it is therefore incumbent upon reformers to address these specific conceptualizations frankly, explicitly, and proactively, for no resource-intensive program of qualitative assessment can succeed otherwise.

In the particular case of community colleges – and now, extended perhaps to the for-profits – where faculty carry heavier (often much heavier) teaching loads in exchange for a release from research obligations, the single most valued resource instructors have is their *autonomy* (Grubb, Worthen, Byrd, Webb, Badway, Case, Goto, & Villeneuve, 1999; Tagg, 2012). Even in cases of standardized curricula that are handed down to instructors who then function as little more than tutors and graders in practice, the freedom of self-determination in

instructional delivery forms a core value to these non-tenure-track faculty. Indeed, the more locked-down and standardized a pedagogy becomes, the more preciously each small remaining degree of academic freedom will be clung to, and each new encroachment is likely to be met with increasingly-determined resistance, as per Darling-Hammond (1997) above.

In the particular case of adjuncts and other contingent faculty, leaders need to weigh the increased inclination toward compliance with seemingly-capricious administrative directives that can typify adjunct working conditions against the comparatively high turnover rates these faculty members are prone to (from issues both inside and outside the institution's control, which is a separate topic entirely). Although it can be easier to implement an ongoing assessment program when relying principally upon faculty who are less disposed than their tenure-track colleagues to stubbornly defend a *status quo* – since the extant culture barely includes those adjuncts in the first place – the accompanying high turnover rate among such faculty creates problems in itself for developing and maintaining the validity of qualitative assessment technologies. Without longitudinal participation from a steady quorum of expert faculty who themselves are directly involved in a representative sample of the instructional delivery, each new round of program assessment will essentially be starting over from scratch, and in the case of content-validity-contingent formative and summative qualitative assessment, this may be an especially formidable obstacle to effectiveness as well as a major cost burden (Katz, 2010).

For administrative leaders, the key is to find ways to improve teaching and learning that do not significantly impact the traditional path to tenure and promotion. Such innovations must always be couched in terms that the faculty will perceive as not imposing a net loss upon what is professionally valuable to them.

Involving the Faculty in the Process of Change

Miller (2012) notes that there has been "a lot of backsliding" (p. 8) in higher education's acceptance of the need to assess and improve teaching and learning. She notes that the key to the process is fostering "the curiosity of the faculty about the effects of their teaching" (p. 8).

Central to fostering faculty involvement in any change process is to establish *ownership* of the process among them, for any change (from any source, external or internal) must have this *authenticity* in order to be accepted (Ewell, 1984; Senge, 1990, Tagg, 2012). Among leadership gurus, an apocryphally-sourced quote attributed to 20th Century author Antoine de Saint-Exupéry is frequently invoked to illuminate this cornerstone notion of modern leadership theory: "If you want to build ship, don't drum up people to collect wood and don't assign them tasks and work, but rather teach them to long for the endless immensity of the sea."

Whereas it is perhaps overreaching to expect higher education faculty to ever "long" for the emergence and growth of an institutional assessment culture that is dedicated to the improvement of postsecondary teaching and learning, the essential, cornerstone question remains: How does one foster and promote faculty involvement in, and ownership of, qualitative assessment?

Tagg (2012) calls for an opening-up of the traditional faculty *endowments* – the vested interests that serve as repositories of resources and value. The traditional linking of hiring, promotion, and tenure decisions to disciplinary research needs to be, if not broken, at least augmented with alternative linkages that do not undervalue teaching and learning and therefore do not deter the faculty embrace of change. In addition, linking these same faculty endowments to collaborative work instead of individual work will enable the faculty to perceive such work as a gain in their respective endowments instead of a loss. Tagg also recommends that this

collaboration be as large as possible, so that faculty involvement in the change process may be as widespread and extensive as possible. To facilitate all of this, it is helpful to establish channels *outside of departments* that allow faculty to build their endowments; the reasoning being that modern institutional departments reinforce specialized research and individual autonomy, but if these are the only paths for faculty development and endowment-building, they drive rigidity and resistance to change.

All this sounds simple and straightforward enough, but the specifics are often elusive, and there are other factors that these optimistic prescriptions for change seem to ignore. Foremost among these is *The Peter Principle* (Peter & Hull, 1969). Simply stated, The Peter Principle is the descriptive phenomenon that in a bureaucratic hierarchy, employees tend to rise to the level of their incompetence. In practice, this means that employees – administrators in particular – are typically promoted and assigned new responsibilities within an institution based on their performance at the previous level of the hierarchy, rather than based on a demonstrated aptitude for the new level of responsibility. Once an employee reaches a job level exceeding that employee's functional capabilities, promotion of the employee ceases, and the employee remains in the hierarchy, attempting to function at a level beyond his or her abilities.

In the years since The Peter Principle was first popularized among management theorists, little has been done to avoid it or alleviate the problems it creates. Indeed, the principle has now been informally extended (by Laurence J. Peter himself) from personnel contexts into to technological ones: The Generalized Peter Principle asserts that any working technology will tend to be applied to progressively more challenging applications until it fails. Unlike personnel scenarios, wherein employees may be retained in unsuitable positions (particularly in situations where their incompetence protects those above them from having their own incompetence

revealed), technology that is inadequate to a new application may often be discarded, rather than simply being retained while remaining ineffective.

To address this from the perspective of qualitative assessment program leadership, leaders must confront The Peter Principle in both its forms. From the human resources side, although it is often preferable in non-postsecondary-education context to provide comfortably-competent employees with salary increases instead of promotions, this is seldom suitable for rank-conscious higher education faculty. The alternative solution then becomes essential: employees must receive extensive training for new responsibilities prior to taking on those responsibilities, so that any emergent incompetencies may be identified before the employee (faculty member) is allowed to take on such new responsibilities. This crucial point is not generally addressed in the advice to be found in the literature, as it clearly will increase the resource costs necessary for implementing an assessment program, and yet, it appears to me to be as essential to planning as any other item – arguably even more so if leaders want to facilitate engagement with and ownership of the resource-intensive qualitative assessment process. As faculty may often lack experience with qualitative assessment, it is vital to any program that faculty be provided with the opportunity to first understand the technology, so that they may consequently best consider how to implement, use, and draw inferences from it.

One of the most-often reported benefits of the assessment projects we run in the Ole Miss Center for Writing and Rhetoric is giving hands-on, practical familiarity with using rubrics to our graduate instructors. These instructors may have been exposed to rubrics when their own work was previously graded using them, and they are guided to use rubrics in assessing their own students' learning, but these novice instructors are frequently impressed by how rubrics may also be used in the overall program assessment we engage in regularly, and participation in that

process expands and extends their understanding of the technology considerably. In effect, we provide training in rubric use for program assessment that has benefits to our individual instructors' student assessments. Once we can persuade (with modest financial incentives) our graduate students to invest the time in our program assessment, they are usually very interested in participating in any future program assessment projects that are subsequently announced – not merely for the small amount of extra money in their paychecks, but for the professional development opportunity it provides.

The technological problem arising from the Generalized Peter Principle is simpler to resolve than the more-basic, personnel-based one. Maintaining the validity of a qualitative assessment technology (a rubric and its inferences) is already built in to the process of evaluating and revising the rubric, and therefore it is a simple matter to evolve a rubric (or a rubric-based program or institutional assessment) to meet any additional requirements or to examine additional outcomes. The technology of rubric-based assessment never becomes obsolete.

When attempting to implement the leadership recommendations I have set out above, it will be helpful for leaders to additionally bear in mind two opposing, yet compatible, models to guide specific decision-making. The first of these is *General Systems Theory* (Von Bertalanffy, 1972). It is essential that leaders consider assessment programs from the perspective of systems theory: a program should always be considered as an interconnected series of parts, each one of which may affect any other part or even the functioning of the system as a whole. On the one hand, this is the basis for the modern notion of a *learning organization* (Senge, 1990) which seeks to leverage the skills and abilities of an institution's members between and across the divisions of the entire organization in an effort to maximize outcomes, but on the other hand, it means that every decision to be implemented is likely to have repercussions anywhere and

everywhere else within the project and the institution, and must therefore be understood in terms of its place within the broader context. Although this task may be non-trivial, it is my aim to have provided, in the previous chapters, the basis for both contextual and theoretical frameworks to facilitate this analysis.

Foremost among these is the other principle that I recommend be used to guide decision-making: *subsidiarity*. Subsidiarity, especially as incorporated into European Union law, is the principle that any organizational matter is best handled by the lowest, least-centralized level of the organizational hierarchy that can effectively address that particular matter. As a demonstrated precondition for effective qualitative assessment, bottom-up methodologies are required to establish the content validity of rubric-based technologies of assessment. As a prerequisite for program success, faculty engagement with and ownership of the process of assessment – particularly resource-intensive qualitative assessment – is likewise crucial. The piece of advice to administrators that I find missing in the literature is that pushing authority for decision-making as far down the institutional hierarchy as it can functionally go is a tacit requirement for program success, and is therefore and essential administrative philosophy. Not only does this reinforce the tradition of shared governance when applied to higher education institutions, but it can guide administrative processes beyond outcomes assessment.

Metarubrics

Finally, administrators should never lose sight of the fact that rubrics may be employed at all levels of an assessment program, and may even be used to assess the effectiveness of other rubrics. Using the methodological, conceptual, and managerial frameworks I have set out, a rubric may be designed, used, and validated for any definable outcome at any institutional level. The more pervasive rubric use becomes, not only inside the classroom, but at all levels of the

institution, the more opportunities to understand meaningful higher education outcomes will be created.

Final Perspective

Ultimately then, effective qualitative assessment is the result of effective rubrics being combined with effective leadership and program management, and the importance of each of these elements to the overall success of a qualitative assessment program cannot be understated. Overall, for professional and intuitional development purposes, I have sought in this research to solve the mystery of why qualitative assessment technologies are not as widespread as a postsecondary educator might naïvely expect them to be, and thereby gain some insight into rectifying the situation in a manner that improves educational delivery. In so doing, I have identified what are, at least to me and my perspective yet hopefully to others' as well, useful and informative frameworks – both practical and theoretical – that have already demonstrated their utility to the administrative challenges I have found myself grappling with regularly in support of program assessment, especially as it is pertinent to institutional accreditation concerns.

Unguided research of the type upon which I have relied heavily in this study can itself be a highly resource-intensive process, and I am fortunate to have been able to negotiate my way into an opportunity to pursue my investigations at length. Going forward, it is now incumbent upon me to promote and extend my findings. Rubrics are the easy part; I have found that the mechanisms of effective rubric use are well-established, but the *knowledge* of how to use them effectively is not, and therefore the need for dissemination of that knowledge is clear to me. It is the leadership challenges that are the hard part of qualitative assessment; promoting and facilitating faculty engagement with, and ownership of, the technologies of qualitative assessment is a major task, but now that I have constructed at least a rudimentary map of the

terrain, further research into how to most effectively accomplish this objective should have a reliable model and direction upon which to build.

LIST OF REFERENCES

References

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Arendt, H. (1998). *The human condition, 2nd ed.* Chicago, IL: University of Chicago Press.

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions.* New York, NY: HarperCollins.

Asher, N. (2009). Considering curriculum questions and the public good in the postcolonial, global, 21st-Century context. *Curriculum Inquiry, 39*(1), 193-204.

Auerbach, D. (2012). The stupidity of computers. *n+1, Issue 13*, np. Retrieved from http://nplusonemag.com/the-stupidity-of-computers

Baker, B. (2004). The functional liminality of the not-dead-yet-students, or, how public schooling became compulsory: A history. *Rethinking History, 8*(1), 5-49.

Banta, T. W. (2006). Reliving the history of large-scale assessment in higher education. *Assessment Update, 18*(4), 3-15. 3p.

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22-27.

Bourdieu, P. (1977). *Outline of a theory of practice.* (R. Nice, Trans.). Cambridge, UK: Cambridge University Press.

Broadhead, P. (2002). The making of a curriculum: How history, politics, and personal perspectives shape emerging policy and practice. *Scandinavian Journal of Educational Research, 46*(10), 47-64.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. (2000). What is construct validity?. *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(2), 8-12.

Carless, D. (2005). Prospects for the implementation of assessment for learning. *Assessment in Education, 12*(1), 39-54.

Chun, M. (2010). Taking teaching to (performance) task: Linking pedagogical and assessment practices. *Change: The Magazine of Higher Learning, 42*(2), 22-29.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp.1-32). San Diego, CA: Academic Press.

Cooper, B. (2012). Universities have been taken over by administrators. *The Vancouver Sun.* Retrieved from http://www.vancouversun.com/Universities+have+been+taken+over+administrators/6626895/story.html

Creswell, J. W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Cronbach, L. J. (1969). Validation of educational measures. In *Proceedings of the 1969 Invitational Conference on Testing Problems: Toward a theory of achievement measurement.* Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., 443-507). Washington, DC: American Council on Education.

Darling-Hammond, L. (1997). *The right to learn.* San Francisco, CA: Jossey-Bass.

DeAngelis, T. (2003). Why we overestimate our competence. *Monitor on Psychology, 34*(2), 60.

Depaepe, N. (2007, February). Philosophy and history of education: Time to bridge the gap?. *Educational Philosophy & Theory, 39*(1), 28-43.

Dewey, J. (1900). *The school and society (1980 ed.).* London: Southern Illinois University Press.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education (2007 ed.).* New York, NY: Free Press.

Eisenhower, D. D. (1961). *Military-Industrial Complex Speech.* Retrieved from http://www.militaryindustrialcomplex.com/eisenhower_farewell_address.asp

Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen, & R. Punämaki (Eds.), *Perspectives on activity theory* (pp. 19-38). Cambridge, UK: Cambridge University Press.

Ewell, P. T. (1984). *The self-regarding institution: Information for excellence.* Boulder, CO: National Center for Higher Education Management.

Ewell, P. T., & Jones, D. P. (1985). *The costs of assessment.* Boulder, CO: National Center for Higher Education Management.

Fairweather, J. S. (1996). *Faculty work and public trust: Restoring the value of teaching and public service in American academic life.* Boston, MA: Allyn and Bacon.

Fendler, L., & Muzaffar, I. (2008). The history of the bell curve: Sorting and the idea of normal. *Educational Theory, 58*(1).

Foucault, M. (1972). *Archaeology of knowledge.* New York, NY: Pantheon.

Foucault, M. (1994). Le sujet et le pouvoir [The subject and the power]. In Defert, D., Ewald, F., & Lagrange, J. (Eds.), *Dits et écrites [Said and written] IV 1980-1988* (231-232). Paris, FRA: Gallimard.

Freire, P. (2006). *Pedagogy of the oppressed, 30th anniversary ed.* New York: Continuum.

Fulbright, W. J. (1970). The war and its effects: The military-industrial-academic complex. In Schiller, H. I. (Ed.), *Super-State: Readings in the Military-Industrial Complex.* Urbana, IL: University of Illinois.

Gall, M. D., Gall, J. P., & Borg, W. R. (2006). *Educational research* (8th ed.). Boston, MA: Allyn and Bacon.

Gallagher, C. W. (2011). Being there: (Re)making the assessment scene. *College Composition and Communication, 62*(3), 450-476.

Gardner, H. (1995, Winter). Cracking open the IQ box. *The American Prospect, 20,* 71-80.

Gilbert, G. (2010). Making faculty count in higher education assessment. *Academe, 96*(5), 25-27.

Giroux, H. A. (1983). *Critical Theory & Educational Practice.* Australia: Deakin University Press.

Giroux, H. A. (1990). *Curriculum Discourse as Postmodernist Critical Practice.* Australia: Deakin University Press.

Giroux, H. A. (1997) *Pedagogy and the politics of hope: Theory, culture, and schooling, a critical reader.* Boulder, CO: Westview Press.

Giroux, H. A. (2003). *The abandoned generation: Democracy beyond the culture of fear.* New York, NY: Palgrave Macmillan.

Giroux, H. A. (2005). *Schooling and the struggle for public life: Democracy's promise and education's challenge (Cultural politics and the promise of democracy).* Boulder, CO: Paradigm Publishers.

Giroux, H. A. (2007). *The university in chains: Confronting the military-industrial-academic complex.* Boulder, CO: Paradigm Publishers.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine Publishing Company.

Goodhart, C. A. E. (1975). Monetary relationships: A view from Threadneedle Street. *Papers in Monetary Economics (Reserve Bank of Australia), 1*.

Gould, S. J. (1981). *The mismeasure of man (1996 ed.).* New York, NY: W. W. Norton & Company, Inc.

Griffiths, J., Vidovich, L., & Chapman, A. (2008). Outcomes approaches to assessment: Comparing non-government and government case-study schools in Western Australia. *The Curriculum Journal, 19*(3), 161-175.

Grubb, W. N., Worthen, H., Byrd, B., Webb, E., Badway, N., Case, C., Goto, S., & Villeneuve, J. C. (1999). *Honored but invisible: An inside look at teaching in community colleges.* New York, NY: Routledge.

Hacking, I. (1990). *The taming of chance.* Cambridge, UK: Cambridge University Press.

Hacking, I. (2002). *Historical ontology.* Cambridge, MA: Harvard University Press.

Harris, J. (1986). Assessing outcomes in higher education. In C. Adelman (Ed.), *Assessment in American higher education: Issues and contexts.* Washington, DC: Office of Educational Research and Improvement.

Helfenbein, R. J., & Shudak, N. J. (2009). Reconstructing/reimagining democratic education: From context to theory to practice. *Educational Studies, 45,* 5-23.

Herrington, A., & Moran, C. (Eds.). (1992). *Writing, teaching, and learning in the disciplines.* New York, NY: Modern Language Association.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York, NY: Free Press.

Hersh, R. H., & Merrow, J. (Eds.). (2005). *Declining by degrees: Higher education at risk.* New York, NY: Palgrave Macmillan.

Holt, J. (2012). *Why does the world exist?: An existential detective story.* New York, NY: Liveright (Norton).

Hopmann, S. T. (2008). No child, no school, no state left behind: Schooling in the age of accountability. *Journal of Curriculum Studies, 40*(4), 417-456.

Huber, M., & Hutchings, P. (2005). *The advancement of learning: Building the teaching commons.* San Francisco, CA: Jossey-Bass.

Hutchings, P. (2010). Opening doors to faculty involvement in assessment. Retrieved from http://www.learningoutcomeassessment.org/documents/PatHutchings_000.pdf

Jones, E. A., Voorhees, R. A., & Paulson, K. (2002). *Defining and assessing learning: Exploring competency-based initiatives (NCES 2002-159).* Washington, DC: United States Department of Education, National Center for Education Statistics.

Kahneman, D., & Tversky, A. (2000). Choices, values, and frames. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 1-16). New York, NY: Russell Sage Foundation/Cambridge University Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin, 112,* 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education/Praeger Publishers.

Katz, S. (2010). Beyond crude measurement and consumerism. *Academe, 96*(5), 16-20.

Krippendorf, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61-70.

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology.* Thousand Oaks, CA: Sage.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Lave, J. (1993). The practice of learning. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 3-32). Cambridge, UK: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Lewis, T. (2007). Biopolitical utopianism in educational theory. *Educational Philosophy and Theory, 39*(7), 683-702.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437-448.

López, C. L. (2002). Assessment of student learning: Challenges and strategies. *The Journal of Academic Librarianship, 28,* 356-367.

Margonis, F. (2009). John Dewey's radicalized visions of the student and classroom community. *Educational Theory, 59*(1), 17-39.

Mason, M. (2008). Complexity theory and the philosophy of education. *Educational Philosophy and Theory, 40*(1), 4-18.

Masschelein, J. (2004). How to conceive of critical educational theory today?. *Journal of Philosophy of Education, 38*(3), 351-365.

Masschelein, J., & Quaghebeur, K. (2005). Participation for better or for worse?. *Journal of Philosophy of Education, 39*(1), 51-65.

Messick, S. (1980). Test validity and the ethics of measurement. *American Psychologist, 35*, 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35-44.

Miller, M. A. (2012). From denial to acceptance: The stages of assessment. Retrieved from http://www.learningoutcomeassessment.org/documents/Miller.pdf

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*, 470-476.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Chapter 4: Validity in educational assessment. *Review of Research in Education 2006, 30*(1), 109-162.

Murray, C. (1995, May). 'The Bell Curve' and its critics. *Commentary,* p.23.

National Institute for Learning Outcomes Assessment. (2012). *From denial to acceptance: The stages of assessment.* Champaign, IL: M. A. Miller.

Neisser, U., & Boodoo, G. (1996, February). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77.

Nietzsche, F. (1887). *On the genealogy of morals (Smith, D., Trans., 1996 repr.).* London, UK: Oxford University Press.

Nkosana, L. (2008). Attitudinal obstacles to curriculum and assessment reform. *Language Teaching Research, 12*(2), 287-312.

Nussbaum, M. C. (2005). Citizens of the World. In L. R. Lattuca, J. G. Haworth, & C. E. Conrad (Eds.), *College and University Curriculum: Developing and Cultivating Programs of Study that Enhance Student Learning* (pp. 242-261). United States: Pearson Custom Publishing.

Organisation for Economic Co-operation and Development [OECD]. 2010. *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)*. OECD Publishing. http://dx.doi.org/10.1787/9789264091559-en

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Peter, L. J., & Hull, R. (1969). *The Peter Principle: Why things always go wrong.* New York, NY: William Morrow and Company.

Peterson, M. W., Einarson, M. K., Augustine, C. H., & Vaughan, D. S. (1999). *Institutional support for student assessment: Methodology and results of a national survey.* Stanford, CA: National Center for Postsecondary Improvement.

Peshkin, A. (2000). The nature of interpretation in qualitative research. *Educational Researcher, 29*(9), 5-9.

Piro, J. M. (2008). Foucault and the architecture of surveillance: Creating regimes of power in schools, shrines, and society. *Educational Studies, 44,* 30-46.

Plucker, J. A. (Ed.) (2003). *Human intelligence: Historical influences, current controversies, teaching resources.* Retrieved from http://www.indiana.edu/~intell

Popkewitz, T. S. (1996). Rethinking decentralization and state/civil society distinctions: The state as a problematic of governing. *Journal Of Education Policy 11*(1), 27-51.

Popper, K. (1959). *The logic of scientific discovery.* (K. Popper, Trans.). London, UK: Routledge. (Original work published in German in 1934 as *Logik der Forschung*).

Reed, T. E., Levin, J., & Malandra, G. H. (2001). Closing the assessment loop by design. *Change: The Magazine of Higher Learning, 43*(5), 44-55.

Rose, N. (1996). Governing 'advanced' liberal democracies. In Barry, A., Osborne, T., & Rose, N. (Eds.), *Foucault and political reason: Liberalism, neo-liberalism and rationalities of government* (37-64). London, UK: UCL Press.

Rossman, J. E., & El-Khawas, E. (1987). *Thinking about assessment: Perspectives for presidents and chief academic officers.* Washington, DC: American Council on Education and the American Association for Higher Education.

Ruhe, V. (2002). Issues in the validation of assessment in technology-based distance and distributed learning: What can we learn from Messick's framework?. *International Journal of Testing, 2*(2), 143-159.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision-making. *Journal of Risk and Uncertainty, 1,* 7-59.

Searle, J. (1990). The storm over the university. *The New York Review of Books, 37*(19), 34-43.

Schuster, J. H., & Finkelstein, M. J. (2006). *The American faculty: He restructuring of academic work and careers.* Baltimore, MD: Johns Hopkins University Press.

Senge, P. (1990). *The fifth discipline: The art & practice of the learning organization* (2006 ed.). United States: Crown Publishing.

Shatzky, J. (2012). The mistrustees. Retrieved from http://academeblog.org/2012/09/14/the-mistrustees/

Shavelson, R. (2008). The Spellings Commission Report and the Collegiate Learning Assessment. In EDUCAUSE (Ed.), *Forum futures 2008* (pp. 35-38). Louisville, CO: EDUCAUSE.

Sheppard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450.

Shipman, D., Aloi, S., & Jones, E. A. (2003). Addressing key challenges in higher education assessment. *The Journal of General Education, 52*(4), 335-346.

Shirow, M. [士郎正宗] (Creator). (2004). *Appleseed* [Motion picture]. Japan: TOHO.

Simons, M., & Masschelein, J. (2006). The learning society and governmentality: An introduction. *Educational Philosophy and Theory, 38*(4), 417-430.

Simons, M., & Masschelein, J. (2008, November). The governmentalization of learning and the assemblage of a learning apparatus. *Educational Theory, 58*(4), 391-415.

Slaney, K. L., & Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31*(2013), 4-12.

Smedley, A. (1998). *AAA statement on race.* Retrieved from http://www.aaanet.org/issues/policy-advocacy/AAA-Statement-on-Race.cfm

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning.* Sterling, VA: Stylus Publishing, LLC.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: The Belknap Press of Harvard University Press.

Stross, R. (2012, June 10). The algorithm didn't like my essay. *The New York Times*, New York edition, p. BU3.

Tagg, J. (2012). Why does the faculty resist change?. *Change: The Magazine of Higher Learning, 44*(1), 6-15.

Tarski, A. (1983). The concept of truth in formalized languages (Woodger, J. H., Trans.). In J. Cocoran (Ed.), *Logic, semantics, mathematics* (pp. 152-278). Indianapolis, IN: Hackett. (Original work published in 1936 in German as Der Wahrheitsbegriff in den formalisierten Sprachen, *Studia Philoshophica, 1*, 261-405.)

Terenzini, P. T. (1989). Assessment with open eyes: Pitfalls in studying student outcomes. *Journal of Higher Education, 60*(6), 644-664.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization, 1,* 39-60.

Thayer-Bacon, B. (2008). Democracies-always-in-the-making: Maxine Greene's influence. *Educational Studies, 44,* 256-269.

Von Bertalanffy, L. (1972). The history and status of general systems theory. *The Academy of Management Journal, 15*(4), 407-426.

Wallace, B., & Graves, W. (1995). *The poisoned apple: The bell-curve crisis and how our schools create mediocrity and failure.* New York, NY: St. Martin's.

Wertsch, J. V. (1998). *Mind as action.* Oxford, UK: Oxford University Press.

Whitson, K. (1998). Key skills and curriculum reform. *Studies in Higher Education, 23*(3), 307-319.

Winerip, M. (2012, April 23). Facing a robo-grader? No Worries. Just keep obfuscating mellifluously. *The New York Times*, New York edition, p. A11.

LIST OF APPENDICIES

APPENDIX A: INFORMATION SHEET AND CONSENT FORM

Appendix A

INFORMATION SHEET AND INTERVIEW CONSENT FORM

Information about a Dissertation Research Study
Title: Improving Qualitative Assessment in Higher Education

Investigator
Chad W. Russell
Department of Leadership and Counselor Education
120 Guyton Hall
The University of Mississippi
(662) 236-4020
crussell@olemiss.edu

Dissertation Chair
Lori A. Wolff, Ph.D., J.D.
Department of Leadership and Counselor Education
139 Guyton Hall
The University of Mississippi
(662) 915-5791
lawolff@olemiss.edu

Description
We seek to identify and understand best practices in qualitative assessment within higher education. Specifically, we want to investigate the professional experiences and perspectives of educators and administrators who report, or are reported to have had, successful and/or valuable experiences in the innovation and/or application of qualitative assessment methods and methodologies in higher education contexts.

In order to investigate this issue, we are inviting you to take part in the research. Your voluntary participation would consist of a guided interview to both describe and explore the meaning of your experience(s) with qualitative assessment. The interview will consist of some initial, pre-written questions to frame your experience within the same general context as that of other interviewees, and then some open-ended follow-up questions to give you an opportunity to provide further details as you may think are relevant. A list of the pre-written interview questions can be provided prior to the interview, if you request it. Ask us, at any point, if there is anything about this research that is not clear or if you would like more information. Take the time to decide whether or not you wish to take part.

Risks and Benefits
You will be interviewed in your professional capacity as an educator and/or an administrator, and therefore may expect and enjoy an appropriate degree of academic freedom in your response. You may be identified not only by name, but also by institutional title and position in the published research, which may be a risk or a benefit, depending on your perspective.

Cost and Payments
The basic interview should last no more than 30 minutes; the length of any open-ended follow-up elaborations will be determined by the mutual consent of both you and the interviewer.

Participation in this study is a purely voluntary service, and will be without particular remuneration.

Confidentiality
You may be identified by name, title, and position, as appropriate, in the published research. The recording of your interview will remain the personal property of the investigator, but the transcript of the interview (annotated, coded, or otherwise) may be published as an appendix to the dissertation. If you request confidentiality at any point or on any issue, the request will be respected, but this may limit the usefulness of your responses to the research.

Right to Withdraw
You do not have to take part in this study. If you start the study and decide that you do not want to finish, all you have to do is to tell Chad Russell (or Lori Wolff) in person, by letter, by e-mail, or by telephone at the Department of Leadership and Counselor Education, 120 Guyton Hall, The University of Mississippi, University MS 38677, or (662) 236-4020. You may withdraw from this study at any time prior to its acceptance for publication at no penalty. The interview data of participants who withdraw will not be published nor shared with third parties, and it will be destroyed at the conclusion of the investigation.

The researchers may terminate your participation in the study without regard to your consent and for any reason, such as protecting your safety and protecting the integrity of the research data.

IRB Approval
This study has been reviewed by The University of Mississippi's Institutional Review Board (IRB). The IRB has determined that this study fulfills the human research subject protections obligations required by state and federal law and University policies. If you have any questions, concerns, or reports regarding your rights as a participant of research, please contact the IRB at (662) 915-7482, regarding Ole Miss IRB protocol 11-120.

Statement of Consent
I have read the above information. I have been given a copy of this form. I have had an opportunity to ask questions, and I have received answers. I consent to participate in the study, and to have the information I provide published as part of the dissertation research.

Signature of Participant Date

APPENDIX B: PROTOCOL FOR INTERVIEW QUESTIONS

Appendix B

Initial Guide for Semi-Structured Interviews — Protocol for Questions

(Designed for use in face-to-face, telephonic, and correspondence interview formats.)

[Begin with noting the time, date, place, and purpose of interview. Identify self and purpose of interview, including specific qualitative assessment method being investigated. Identify and thank interviewee.]

1. For the record and in your own words, please state your name and position/title.

    1a. What is your background in assessment?

2. What can you tell me about [qualitative assessment program/instrument/experience being investigated] in general?

    2a. How did it and/or its use come about?

        2a1. Why was the purpose behind implementing it?

    2b. What was/is your involvement with it?

    2c. What were the results/outcomes of it?

        2c1. Was it a one-time trial, or is it ongoing, and why?

        2c2. Has it been modified or replaced, and why?

    2d. What was/is the value of it?

        2d1. To your institution?

        2d2. To your program/division?

        2d3. To your colleagues?

        2d4. To the students?

3. What is your professional opinion of [program/instrument/experience] as a method of qualitative assessment?

3a. Do your colleagues share this opinion, or not?

3b. What sort of feedback did you receive?

3b1. From colleagues?

3b2. From students?

3b3. From accreditors?

4. What do you think of qualitative assessment both in and of itself, and in contrast to quantitative assessment?

4a. What is the purpose of assessment?

4a1. For educators?

4a2. For institutions?

4a3. For students?

4a4. For other stakeholders?

4b. How would you characterize the difference(s) between qualitative and quantitative assessment in meeting these goals?

5. What was/is the most important thing gained from your experience with [qualitative assessment program/instrument/experience]? [Open-ended; may require follow-up.]

5a. Is there anything else about the history, design, features, and/or outcomes of your experience that you would like to mention or elaborate upon? [Open-ended; may require follow-up.]

6. What recommendations do you have, or did your institution generate, as a result of using this particular qualitative assessment? [Particularly open-ended; may require extensive follow-up.]

7. Is there anything else you would like to add that you think is relevant to the research? [Open-ended; may require follow-up.]

[Thank interviewee, note time.]

APPENDIX C: PERELMAN ESSAY

Appendix C

The following is an essay question response written by Les Perelman and awarded a top grade of 6 by e-Rater, the automated grader developed by ETS. This is believed to be a verbatim transcription of a timed-writing original, and any typographic errors – unintentional or otherwise – are therefore preserved here.

Question: "The rising cost of a college education is the fault of students who demand that colleges offer students luxuries unheard of by earlier generations of college students—single dorm rooms, private bathrooms, gourmet meals, etc."

Discuss the extent to which you agree or disagree with this opinion. Support your views with specific reasons and examples from your own experience, observations, or reading.

Response: In today's society, college is ambiguous. We need it to live, but we also need it to love. Moreover, without college most of the world's learning would be egregious. College, however, has myriad costs. One of the most important issues facing the world is how to reduce college costs. Some have argued that college costs are due to the luxuries students now expect. Others have argued that the costs are a result of athletics. In reality, high college costs are the result of excessive pay for teaching assistants.

I live in a luxury dorm. In reality, it costs no more than rat infested rooms at a Motel Six. The best minds of my generation were destroyed by madness, starving hysterical naked, and publishing obscene odes on the windows of the skull. Luxury dorms pay for themselves because they generate thousand and thousands of dollars of revenue. In the Middle Ages, the University of Paris grew because it provided comfortable accommodations for each of its students, large

rooms with servants and legs of mutton. Although they are expensive, these rooms are necessary to learning. The second reason for the five-paragraph theme is that it makes you focus on a single topic. Some people start writing on the usual topic, like TV commercials, and they wind up all over the place, talking about where TV came from or capitalism or health foods or whatever. But with only five paragraphs and one topic you're not tempted to get beyond your original idea, like commercials are a good source of information about products. You give your three examples, and zap! you're done. This is another way the five-paragraph theme keeps you from thinking too much.

Teaching assistants are paid an excessive amount of money. The average teaching assistant makes six times as much money as college presidents. In addition, they often receive a plethora of extra benefits such as private jets, vacations in the south seas, a staring roles in motion pictures. Moreover, in the Dickens novel Great Expectation, Pip makes his fortune by being a teaching assistant. It doesn't matter what the subject is, since there are three parts to everything you can think of. If you can't think of more than two, you just have to think harder or come up with something that might fit. An example will often work, like the three causes of the Civil War or abortion or reasons why the ridiculous twenty-one-year-old limit for drinking alcohol should be abolished. A worse problem is when you wind up with more than three subtopics, since sometimes you want to talk about all of them.

There are three main reasons while Teaching Assistants receive such high remuneration. First, they have the most powerful union in the United States. Their union is greater than the Teamsters or Freemasons, although it is slightly smaller than the international secret society of the Jedi Knights. Second, most teaching assistants have political connections, from being children of judges and governors to being the brothers and sisters of kings and princes. In Heart

of Darkness, Mr. Kurtz is a teaching assistant because of his connections, and he ruins all the universities that employ him. Finally, teaching assistants are able to exercise mind control over the rest of the university community. The last reason to write this way is the most important. Once you have it down, you can use it for practically anything. Does God exist? Well, you can say yes and give three reasons, or no and give three different reasons. It doesn't really matter. You're sure to get a good grade whatever you pick to put into the formula. And that's the real reason for education, to get those good grades without thinking too much and using up too much time.

In conclusion, as Oscar Wilde said, "I can resist everything except temptation." Luxury dorms are not the problem. The problem is greedy teaching assistants. It gives me an organizational scheme that looks like an essay, it limits my focus to one topic and three subtopics so I don't wander about thinking irrelevant thoughts, and it will be useful for whatever writing I do in any subject.1 I don't know why some teachers seem to dislike it so much. They must have a different idea about education than I do.

By Les Perelman

VITA

Chad W. Russell earned the degree of Bachelor of Science in Mathematical Sciences (Computer Science and Systems Analysis), with a Minor in Physics, from the University of North Carolina at Chapel Hill in 1984. In 1990, he earned the Post-Baccalaureate Certificate of Major in Philosophy, with a Minor in Psychology, from the University of North Carolina at Asheville. Then in 1995, he earned the Master of Arts in Philosophy from the University of Miami at Coral Gables, concentrating in Epistemology and Philosophy of Mind.